Université de Montréal

Optimization Tools for Non-Asymptotic Statistics in Exponential Families

par

Rémi Le Priol

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Décembre, 2021

Université de Montréal
Faculté des arts et des sciences

Cette thèse intitulée:

Optimization Tools for Non-Asymptotic Statistics in Exponential Families

présentée par:

Rémi Le Priol

a été évaluée par un jury composé des personnes suivantes:

Ioannis Mitliagkas,	président-rapporteur
Simon Lacoste-Julien,	directeur de recherche
Yoshua Bengio,	codirecteur
Guillaume Rabusseau,	membre du jury
Nicolas Flammarion,	examinateur externe

Thèse acceptée le: .....

Les théorèmes sont démontrés par ceux qui y croient.

Theorems are proved by those who believe in them.

André Weil

Résumé

Les familles exponentielles est une classe de modèles omniprésente en statistique. D'une part, elle peut modéliser n'importe quel type de données. En fait la plupart des distributions communes en font partie : Gaussiennes, variables catégoriques, Poisson, Gamma, Wishart, Dirichlet. D'autre part elle est à la base des modèles linéaires généralisés (GLM), une classe de modèles fondamentale en machine learning. Enfin les mathématiques qui les sous-tendent sont souvent magnifiques, grâce à leur lien avec la dualité convexe et la transformée de Laplace. L'auteur de cette thèse a fréquemment été motivé par cette beauté. Dans cette thèse, nous faisons trois contributions à l'intersection de l'optimisation et des statistiques, qui tournent toutes autour de la famille exponentielle.

La première contribution adapte et améliore un algorithme d'optimisation à variance réduite appelé ascension des coordonnées duales stochastique (SDCA), pour entraîner une classe particulière de GLM appelée champ aléatoire conditionnel (CRF). Les CRF sont un des piliers de la prédiction structurée. Les CRF étaient connus pour être difficiles à entraîner jusqu'à la découverte des technique d'optimisation à variance réduite. Notre version améliorée de SDCA obtient des performances favorables comparées à l'état de l'art antérieur et actuel.

La deuxième contribution s'intéresse à la découverte causale. Les familles exponentielles sont fréquemment utilisées dans les modèles graphiques, et en particulier dans les modèles graphique causaux. Cette contribution mène l'enquête sur une conjecture spécifique qui a attiré l'attention dans de précédents travaux : les modèles causaux s'adaptent plus rapidement aux perturbations de l'environnement. Nos résultats, obtenus à partir de théorèmes d'optimisation, soutiennent cette hypothèse sous certaines conditions. Mais sous d'autre conditions, nos résultats contredisent cette hypothèse . Cela appelle à une précision de cette hypothèse, ou à une sophistication de notre notion de modèle causal.

La troisième contribution s'intéresse à une propriété fondamentale des familles exponentielles. L'une des propriétés les plus séduisantes des familles exponentielles est la forme close de l'estimateur du maximum de vraisemblance (MLE), ou maximum a posteriori (MAP) pour un choix naturel de prior conjugué. Ces deux estimateurs sont utilisés presque partout, souvent sans même y penser. (Combien de fois calcule-t-on une moyenne et une variance pour des données en cloche sans penser au modèle Gaussien sous-jacent ?) Pourtant la littérature actuelle manque de résultats sur la convergence de ces modèles pour des tailles d'échantillons finis, lorsque l'on mesure la qualité de ces modèles avec la divergence de Kullback-Leibler

(KL). Pourtant cette divergence est la mesure de différence standard en théorie de l'information. En établissant un parallèle avec l'optimisation, nous faisons quelques pas vers un tel résultat, et nous relevons quelques directions pouvant mener à des progrès, tant en statistiques qu'en optimisation.

Ces trois contributions mettent des outil d'optimisation au service des statistiques dans les familles exponentielles : améliorer la vitesse d'apprentissage de GLM de prédiction structurée, caractériser la vitesse d'adaptation de modèles causaux, estimer la vitesse d'apprentissage de modèles omniprésents. En traçant des ponts entre statistiques et optimisation, cette thèse fait progresser notre maîtrise de méthodes fondamentales d'apprentissage automatique.

Mots-clés

Apprentissage automatique, famille exponentielle, divergence de Bregman, statistiques non-asymptotiques, taux de convergence, dualité, optimisation stochastique, réduction de variance, prédiction structurée, causalité.

Abstract

Exponential families are a ubiquitous class of models in statistics. On the one hand, they can model any data type. Actually, the most common distributions are exponential families: Gaussians, categorical, Poisson, Gamma, Wishart, or Dirichlet. On the other hand, they sit at the core of generalized linear models (GLM), a foundational class of models in machine learning. They are also supported by beautiful mathematics thanks to their connection with convex duality and the Laplace transform. This beauty is definitely responsible for the existence of this thesis. In this manuscript, we make three contributions at the intersection of optimization and statistics, all revolving around exponential families.

The first contribution adapts and improves a variance reduction optimization algorithm called stochastic dual coordinate ascent (SDCA) to train a particular class of GLM called conditional random fields (CRF). CRF are one of the cornerstones of structured prediction. CRF were notoriously hard to train until the advent of variance reduction techniques, and our improved version of SDCA performs favorably compared to the previous state-of-the-art.

The second contribution focuses on causal discovery. Exponential families are widely used in graphical models, and in particular in causal graphical models. This contribution investigates a specific conjecture that gained some traction in previous work: causal models adapt faster to perturbations of the environment. Using results from optimization, we find strong support for this assumption when the perturbation is coming from an intervention on a cause, and support against this assumption when perturbation is coming from an intervention on an effect. These pieces of evidence are calling for a refinement of the conjecture.

The third contribution addresses a fundamental property of exponential families. One of the most appealing properties of exponential families is its closed-form maximum likelihood estimate (MLE) and maximum a posteriori (MAP) for a natural choice of conjugate prior. These two estimators are used almost everywhere, often unknowingly – how often are mean and variance computed for bell-shaped data without thinking about the Gaussian model they underly? Nevertheless, literature to date lacks results on the finite sample convergence property of the information (Kulback-Leibler) divergence between these estimators and the true distribution. Drawing on a parallel with optimization, we take some steps towards such a result, and we highlight directions for progress both in statistics and optimization.

These three contributions are all using tools from optimization at the service of statistics in exponential families: improving upon an algorithm to learn GLM,

characterizing the adaptation speed of causal models, and estimating the learning speed of ubiquitous models. By tying together optimization and statistics, this thesis is taking a step towards a better understanding of the fundamentals of machine learning.

Keywords

Machine learning, exponential family, Bregman divergence, non-asymptotic statistics, sample complexity, duality, stochastic optimization, variance reduction, structured prediction, causality.

Contents

List of Tables

List of Figures

List of acronyms and abbreviations

AISTATS	conference on Artificial Intelligence and STATisticS
ASG	Average Stochastic Gradient
BCFW	Block-Coordinate Frank-Wolfe
CONLL	conference on COmputational Natural Language Learning
CPU	Central Processing Unit
CRF	Conditional Random Field
DAG	Directed Acyclic Graph
e.g.	*exempli gratia* [for instance]
ERM	Empirical Risk Minimization
GD	Gradient Descent
GLM	Generalized Linear Model
GPU	Graphical Processing Unit
i.e.	*ide est* [that is]
KL	Kullback-Leibler divergence
MAP	Maximum A Posteriori estimate
MD	Mirror Descent
ML	Machine Learning
MLE	Maximum Likelihood Estimate
NER	Named-Entity Recognition
NUS	Non-Uniform Sampling
OCR	Optical Character Recognition
OEG	Online Exponentiated Gradient
POS	Part Of Speech tagging
RAM	Random Access Memory
resp.	respectively
SAG	Stochastic Average Gradient
SCM	Structural Causal Model
SDCA	Stochastic Dual Coordinate Ascent
SGD	Stochastic Gradient Descent
SMD	Stochastic Mirror Descent
SVRG	Stochastic Variance Reduced Gradient
SVM	Support Vector Machine
UAI	conference on Uncertainty in Artificial Intelligence

Remerciements

À ma mère et mon père pour le soutien qu'ils m'ont toujours offert. À mes deux frères pour la continuité, la présence et les ouvertures qu'ils apportent à ma vie.

Merci Simon pour ces presque 5 années de collaborations, pour ton soutien à travers les succès et les doutes, pour ta bienveillance et ton écoute face à mes choix de vie, qui m'ont entre autres mené à rentrer en France pris par un sentiment d'urgence écologique.

Merci Yoshua d'avoir posé les pierres du Mila, et d'y défendre, une recherche ouverte et créative. Ce laboratoire qui a été pour moi à la fois une maison et un lieu d'aventures.

Merci Gabriel d'avoir ouvert la voie à Montréal (ton feu y a fait fondre la neige).

Merci Romain Lopez. Sans toi je n'aurais peut-être pas envoyé le mail qui m'a amené au Mila.

Pour tous mes collaborateurs et amis de Montréal et d'ailleurs, dans l'ordre de ma mémoire: Thomas, Akram, Gauthier, Ahmed, Sébastien, Tristan, Waïss, Reza, Radu, Hadrien, Damien, Frederik, Tom, Hugo, Falco, Ju, Louve, Robin, Nicole, Thiago, Oscar, Léonard, le Black Yak, Jill, Heidi, Byron, Joe, Simon, Aristide, Deanna, Alejandro the monk, Anna.

Pour ceux qui m'aident aujourd'hui à me réinventer : Lise Dargentolle, X-Urgence Écologique, Vincent, Hervé et les châteaux d'acroyoga.

Et surtout merci à Aurélie de m'aider à écrire ces pages et à écrire ma vie.

1 Introduction

Statistics emerged more than a millennium ago, as frequency counting could be used to decipher encrypted messages. Mathematical optimization started being formalized a few centuries ago, as Newton and Gauss designed the first iterative methods to find an optimum, followed by Cauchy et al. (1847) who invented gradient descent.

Seventy years ago, scientists like Turing created digital computers and started dreaming of artificial intelligence: creating machines that could act and learn as humans do, or better. A step towards this goal is enabling computers to interact with the world based on data: images, sound, text, or any arbitrary tabular data. The most straightforward approach to give them this capability is to code explicit rules: if you receive this input, then you should do this. Unfortunately, this approach, known as expert systems or symbolic artificial intelligence, does not scale with the complexity or the quantity of data: imagine coding rules to identify a dog in an image given the raw string of one million pixels. Almost as early as the first computer, Turing (1950) formulated the idea of a program learning new rules from examples (Muggleton, 2014). Several years later, Rosenblatt (1957) introduced the perceptron, an early form of support vector machines doing exactly that. This is the realm of machine learning where our story takes place.

Modern machine learning fits models to data points. It is essentially a new field at the intersection of statistics and optimization, but some of its core ideas are not new. They even predate the emergence of computers. Legendre and Gauss already applied linear regression to predict planetary movements around 1800. The long history of maximum likelihood can be summarized by the name of Fisher, who applied this principle to logistic regression in the 1930s (Stigler, 2007). Stochastic gradient descent (Robbins and Monro, 1951), the workhorse of modern machine learning, was introduced as a root-finding algorithm.

The long history of statistics and optimization have intertwined to form machine learning as we know it today. This thesis covers some of the interactions between these two fields. Abstracting away, our contributions are addressing three challenges

1. training models faster
2. learning causal features that generalize better
3. characterizing the learning speed for a wide variety of models.

By tapping into the connection between statistics and optimization, this thesis attempts to further progress on these three challenges.

Thesis Outline This thesis presents contributions spanning diverse parts of machine learning. We provide Chapter 2 as a reference for readers that are unfamiliar with any of these parts. In particular, we review the two learning setups that we are using: supervised learning and density estimation via maximum likelihood. We then cover some fundamentals of convex optimization: SGD, variance reduction, and convex duality. Finally, we introduce exponential families, graphical models and briefly introduce causal inference for the newcomers.

Chapter 3 is about training faster. The goal is to train conditional random fields (CRF) as fast as possible. CRF are probabilistic models for structured prediction that have long been notoriously hard to train. For this purpose, we adapt SDCA, a well-known optimization algorithm. On the way, we introduce a new sampling scheme for SDCA, and we prove an improved convergence rate. Empirically, SDCA performs on par or better than other variance reduction algorithms.

Chapter 4 is about optimization and causality. It questions a specific assumption: we expect causal models to adapt faster to perturbations of the world. Is this true? We answer this question for bivariate categorical and multivariate normal data by modeling perturbations as interventions in a ground truth causal model and modeling adaptation as optimization.

Chapter 5 is about optimization and elementary statistics. While maximum likelihood estimates are used everywhere, they are most commonly used with exponential families where the solution has a closed-form. A typical frequentist approach would estimate the average risk of this estimator, and a standard measure of risk would be the KL divergence between this estimator and the ground truth. Nevertheless, we are not aware of any general results upper bounding this expected KL. Surprisingly, this statistical problem is connected with the optimization of non-smooth losses. We review the importance of this problem and showcase recent attempts at solving it.

Finally, the last chapter concludes the thesis by summarizing the contributions and outlining some future research directions in the context of this thesis.

2 Background

In this section, we will review elementary building blocks that are pre-requisites to understand all three contributions. First, we are going to review some fundamental principles of machine learning: empirical risk minimization and maximum likelihood estimation, for supervised learning §2.1.1 or density estimation §2.1.2. Ultimately, we want to answer the following question: How can one minimize the empirical risk? To do so, we will cover some parts of the vast topic that is optimization: stochastic gradient descent §2.2.1, variance reduction §2.2.2 and Fenchel duality §2.2.3. Finally, we will describe some useful probabilistic models: exponential families §2.3.1 (which are at the core of this thesis) probabilistic graphical models §2.3.2, and structural causal models §2.3.3.

2.1 Learning from Data

2.1.1 Supervised Learning

Let us assume we have n data points $\mathcal{D} = (z_1, \dots, z_n)$, that decompose as $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, where we call x_i features, and y_i labels. We want to learn some rules to map newly observed features x to their unobserved labels y. The most common approach as of today is to define:

1. a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parametrized by some vector $\theta \in \mathbb{R}^d$. This model will implicitly contain all the rules that we are unable to explicitly write down. For x and y real vectors, the simplest instance of functions are linear models: $f_\theta(x) = \theta^T x$.

2. a prediction loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that will tell use how well or how poorly we are doing. The simplest instance may be $\ell(y_1, y_2) = \|y_1 - y_2\|^2$.

Then we may learn the mapping by solving the following problem:

$$\min_\theta \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) \ . \tag{2.1}$$

This problem simply minimizes the sum of the loss on all data points. This sum is a proxy for the expectation of the loss on the "true" distribution $\boldsymbol{p}$ from which we sampled cD

$$\mathbb{E}_{(x,y)\sim\boldsymbol{p}}\left[\ell(y, f_\theta(x))\right] \ . \tag{2.2}$$

In learning theory, the expectation (2.2) is called either *generalization error*, either *risk function*. That is why problem (2.1) bears the name *empirical risk minimization*, or ERM to keep it short.

In this thesis, we are going to focus on a variant of ERM: *maximum likelihood estimation*. This special case happens when we assume that all data points were sampled independently and identically from some distribution $\boldsymbol{p}(x, y)$ defined on $\mathcal{X} \times \mathcal{Y}$. Then $f_\theta(x_i)$ may return the log-probabilities (or the log-densities) of y given x_i, e.g.

$$f_\theta(x) = -\log \boldsymbol{p}_\theta(y = \cdot \,|\, x) \in \mathbb{R}^{|\mathcal{Y}|} \ , \tag{2.3}$$

and the loss may return the log-likelihood of y_i given x_i, e.g.

$$\ell(y_i, f_\theta(x_i)) = -\log \boldsymbol{p}_\theta(y_i \,|\, x_i) \ . \tag{2.4}$$

Then, if θ^* solves the problem

$$\theta^* = \operatorname*{argmax}_\theta \boldsymbol{p}_\theta(y_1, \ldots, y_n \,|\, x_1, \ldots, x_n) \tag{2.5}$$

$$= \operatorname*{argmin}_\theta \sum_{i=1}^{n} -\log \boldsymbol{p}_\theta(y_i \,|\, x_i) \ , \tag{2.6}$$

we will call it *maximum (conditional) likelihood estimate* or MLE for short. Prior to supervised learning, statisticians have also studied this framework for density estimation.

2.1.2 Density Estimation

If we are not interested in a division of z between features x and labels y, we may still want to learn the probability distribution $\boldsymbol{p}(z) = \boldsymbol{p}(x, y)$. That is the realm of density estimation. Now θ may parametrize a density model $z \mapsto \boldsymbol{p}_\theta(z)$. To solve this problem, maximizing the likelihood of the dataset is again an interesting approach:

$$\theta^* = \operatorname*{argmin}_\theta \sum_{i=1}^{n} -\log \boldsymbol{p}_\theta(x_i) \ . \tag{2.7}$$

In this case as well, θ^* is the *maximum likelihood estimate* (MLE).

Example 1: Isotropic Gaussian. If the data is made of real vectors $z \in \mathbb{R}^d$, then we may chose to fit an isotropic Gaussian $\boldsymbol{p}_\theta = \mathcal{N}(\mu, \sigma^2 \boldsymbol{I})$. In this case, MLE has a closed form solution with the empirical mean and the empirical variance as estimate for the mean μ and variance σ^2 parameters

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \mu\|^2 \,. \tag{2.8}$$

Example 2: Categorical. If the data is made of categories, e.g., $z \in \{1, \dots, K\}$, then the simplest model is the Categorical distribution $\boldsymbol{p}_\theta = (p_1, \dots, p_K) \in \boldsymbol{\Delta}_K$, which assigns a probability p_k to each label k. We use $\boldsymbol{\Delta}_K$ to denote the K-simplex, e.g., the set of real positive vectors of dimension K that sum to 1:

$$\boldsymbol{\Delta}_K = \left\{ \boldsymbol{p} \in \mathbb{R}^K | \forall k, p_k \geq 0; \sum_{k=1}^{K} p_k = 1 \right\} \,. \tag{2.9}$$

The maximum likelihood estimate of this model counts the occurrence of each label in the data and takes the empirical frequency of each label

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i = k\} \tag{2.10}$$

where the indicator function $\mathbf{1}\{b\}$ is 1 if b is true and 0 otherwise.

MAP. In some situations, we may take a Bayesian stance and assume the true parameter θ is itself a random variable sampled from a known *prior* distribution $\boldsymbol{p}(\theta)$. Combining this prior with our model, we obtain a joint distribution on data and parameters

$$\boldsymbol{p}(z, \theta) := \boldsymbol{p}(z \,|\, \theta)\boldsymbol{p}(z) := \boldsymbol{p}_\theta(z)\boldsymbol{p}(\theta) \,. \tag{2.11}$$

By Bayes rule, we also obtain a posterior distribution on parameters

$$\boldsymbol{p}(\theta \,|\, z) = \frac{\boldsymbol{p}(z, \theta)}{\boldsymbol{p}(z)} \propto \boldsymbol{p}_\theta(z)\boldsymbol{p}(\theta) \tag{2.12}$$

The maximum a posteriori or MAP estimate is then defined as the parameter with maximal posterior density (or posterior mass for discrete parameters)

$$\theta^*_{\text{MAP}} := \operatorname{argmax} \boldsymbol{p}(\theta \,|\, z) \tag{2.13}$$

$$= \operatorname{argmin} - \log \boldsymbol{p}_\theta(z) - \log \boldsymbol{p}(\theta) \tag{2.14}$$

Letting go of the Bayesian perspective, this last minimization problem may be seen as an instance of regularized empirical risk minimization, where the negative log-likelihood of the prior $-\log \boldsymbol{p}(\theta)$ plays the role of the regularizer.

Conjugate Priors. Given some models such as exponential families, there exist families of prior distributions $\mathcal{F} = \{\boldsymbol{p}_\eta(\theta) \mid \eta\}$ such that the posterior distribution also belongs to $\mathcal{F}$, e.g. $\forall \eta, \exists \eta', \boldsymbol{p}_\eta(\theta \mid z) = \boldsymbol{p}_{\eta'}(\theta)$. We refer to such families $\mathcal{F}$ as *conjugate priors*. In our third contribution §5.2 we will cover in detail a generic instance of conjugate priors for exponential families.

Other models. Both Gaussians and categorical models belong to the general class of exponential families, which we will introduce in §2.3.1. These families have a limited capacity: there are some set of distributions that cannot be fit by an exponential family, such as mixture models. There exist much more powerful models such as *normalizing flows* (Rezende and Mohamed, 2015) which are based on neural networks. Normalizing flows also rely on maximizing the likelihood of a dataset. They can model almost any smooth low dimensional density, but their capacity is limited in higher dimensions (Kong and Chaudhuri, 2020).

Beyond MLE. To learn from unlabeled data, there exist many competing approaches to MLE. If we model the data with some unobserved variables, then we enter the realm of variational inference, with powerful algorithms such as variational auto-encoders (Kingma and Welling, 2013). If we are more interested in creating new realistic samples from $\boldsymbol{p}(z)$, then adversarial training may be relevant (Goodfellow et al., 2014). Finally, for training large models, self-supervised learning has recently emerged as the leading set of techniques to learn powerful features, most notably for natural languages (Peters et al., 2018; Devlin et al., 2018).

2.2 Convex Optimization

Convex optimization is the field of mathematics interested in solving problems of the form

$$\min_{\theta \in \Theta} f(\theta) \tag{2.15}$$

where f is a convex real valued function $f : \Theta \to \mathbb{R}$ called the loss or the *objective function*, and Θ is a convex set called the *constraint set*. In this work, we always assume that the problem is unconstrained, i.e., $\Theta = \mathbb{R}^d$, but with an objective function taking possibly infinite values, i.e., $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. The objective is thus implicitly defining a constraint set via its domain $\operatorname{Dom} f = \{\theta \mid f(\theta) < +\infty\}$. For instance, we may encounter $\operatorname{Dom} f = \mathcal{S}_d^+$, the set of symmetric positive definite matrices of order d. Note that this set is open, and consequently, we cannot project onto it. We also assume that f is differentiable.

Gradient Descent. When $\Theta = \mathbb{R}^d$, and if we can compute derivatives of f, the most well-known algorithm to solve this problem is gradient descent. Starting from a random point θ_0, iteratively nudge the parameters in the direction opposite to the gradient of the loss, i.e.,

$$\theta_{t+1} = \theta_t - \gamma_t \nabla f(\theta_t), \tag{GD}$$

where the hyper parameter γ_t is known as the *step-size* or the learning rate. The step-size may be constant, follow a predefined schedule, be found via a line-search, or be adaptive w.r.t. to the past trajectory.

Convergence Analysis. Gradient descent does not converge all the time. We need assumptions on the objective function. Perhaps the most common assumption is smoothness.

Proposition 2.2.1 (smoothness)**.** *A function $f : \Theta \to \mathbb{R}$ is said to be L-smooth if it is differentiable and its gradient is L-Lipschitz, i.e.,*

$$\forall \theta, \nu \in \Theta, \|\nabla f(\theta) - \nabla f(\nu)\| \leq L\|\theta - \nu\| \ . \tag{2.16}$$

If the loss f is convex and L-smooth then gradient descent with constant step-size $\gamma_t = \frac{1}{L}$ converges to a minimum θ^* at a rate $O(\frac{1}{t})$ (Nesterov, 2004a, corollary 2.1.2). Smoothness has a sibling assumption: strong-convexity.

Definition 2.2.2 (strong-convexity)**.** *A function $f : \Theta \to \mathbb{R}$ is said to beμ-strongly convex if $\theta \mapsto f(\theta) - \frac{L}{2}\|\theta\|^2$ is convex.*

If f is both μ-smooth and L-strongly convex, then gradient descent with constant step-size $\gamma_t = \frac{2}{\mu + L}$ converges at a *linear rate* $O(e^{-\frac{t}{\kappa}})$ where $\kappa = \frac{L}{\mu} \geq 1$ is known as the condition number of the problem (Nesterov, 2004a, theorem 2.1.15).

Self-concordance. In contributions 2 and 3, we will face the log-likelihood of a multivariate normal variable. This objective includes a term $g(\Lambda) = -\log\det\Lambda$ where $\Lambda \in \mathcal{S}_n^+$ is the positive definite precision matrix. The objective g shoots up to $+\infty$ when Λ gets eigenvalues close from zero. This means that its gradient is not a Lipschitz function. In fact g is neither smooth nor strongly convex. This kind of log-barrier objectives often comes up in interior point methods for solving constrained convex problems. To analyze Newton's method applied on these objectives, a new assumption upper bounding the third derivative with the second derivative was introduced (Nemirosky and Yudin, 1983).

Definition 2.2.3 (self-concordance)**.** *(Nesterov, 2004a, definition 4.1.1) A convex function $f : \Theta \to \mathbb{R}$ is said to be self-concordant if*

$$\forall \theta \in \Theta, \forall u, \nabla^3 f(\theta)[u, u, u] \leq 2\nabla^2 f(\theta)[u, u]^{\frac{3}{2}} \tag{2.17}$$

where we evaluated the third-order tensor $\nabla^3 f(\theta)$ in u, u, u.

Self-concordance was popularized by making the analysis Newton's method affine invariant. Indeed Newton's method is affine invariant, meaning that any affine re-parametrization of the problem does not modify the algorithm. And yet all analysis were stuck with Lispchitz constants which are *not* affine invariant. Self-concordance overcame this fundamental limitation. It was then popularized in machine learning by Bach (2010) who showed its usefulness on logistic regression. It used a fundamental consequence of self-concordance: the objective can be sandwiched between quadratics in a clear neighborhood around its optimum. In this thesis, we are interested in self-concordance because many exponential families have self-concordant log-likelihood. We explore this fact in our third contribution.

2.2.1 Stochastic Gradient Descent

When f has some structure, it is possible to design more efficient algorithms than gradient descent. As seen in Eqs. (2.6) and (2.7), machine learning is generally interested in minimizing an expected loss over a dataset, i.e.,

$$\min_{\theta} F(\theta) := \frac{1}{n} \sum_{i=1}^{n} f(\theta, x_i) \; . \tag{2.18}$$

For instance, this loss may be the negative log-likelihood $f(\theta, z_i) = -\log \boldsymbol{p}_\theta(z_i)$. The gradient of this empirical loss is the sum of gradients on each data points as follow,

$$\nabla F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f(\theta, x_i) \; . \tag{2.19}$$

Modern datasets are huge. They often contain millions, if not billions, of high dimensional data points such as images. As a consequence exact minimization is no longer the bottleneck in learning (Bottou and Bousquet, 2008). Computing the exact gradient, a sum with a billion terms, is no longer affordable. Instead, it is much more efficient to compute gradients for a few data point at a time, and take a step in their opposite direction in the hope of minimizing the loss

$$\theta_{t+1} = \theta_t - \gamma_t \nabla_\theta f(\theta, x_i) \tag{SGD}$$

where i is sampled uniformly from $\{1, \dots, n\}$. This is *stochastic gradient descent*[1] (SGD). It was first devised by Robbins and Monro (1951) to find the zeros of a stochastic function.

[1] Contrary to gradient descent, SGD is not guaranteed to decrease the objective value at every step. As such, it is not a descent algorithm. We should rigorously call it the stochastic gradient method, but SGD has become the standard acronym in the community; therefore, we will stick with it.

A special case happens when we sample each data point only once. Then SGD minimizes the true population risk

$$\min_{\theta} \; F(\theta) := \mathbb{E}_{x\sim \boldsymbol{p}} \left[f(\theta, x) \right] \; . \tag{2.20}$$

In our second contribution, we use this fact by interpreting convergence rates of SGD as a bound on the sample complexity of the model.

Convergence Analysis. Let us review a simplification of the modern convergence analysis from Gower et al. (2019). Assume that

- $\forall x, f(\cdot, x)$ is L-smooth,
- F is strongly convex, minimized by θ^*,
- the gradient noise at the optimum is finite, i.e.,

$$\sigma^2 := \mathbb{E}_{x\sim \boldsymbol{p}} \left[\|\nabla f(\theta^*, x)\|^2 \right] < \infty.$$

Then iterates of SGD with constant step-size step size $\gamma_t = \gamma \in (0, \frac{1}{2L}]$ verify (Gower et al., 2019, theorem 3.1)

$$\mathbb{E}\left[F(\theta_t)\right] - F(\theta^*) \leq \frac{L}{2}(1-\gamma\mu)^t \|\theta_0 - \theta^*\|^2 + \gamma\sigma^2 \frac{L}{\mu}, \tag{2.21}$$

where the expectation is taken over the stochastic procedure. In other words, SGD with constant step-size converges at a linear rate to a variance ball around the optimum, and the size of this variance ball is proportional to the step-size γ, the gradient noise at the optimum σ^2 and the condition number $\frac{L}{\mu}$. To overcome this variance ball issue, we may progressively decrease the learning rate $\gamma_t \in O(\frac{1}{t})$ to obtain a convergence rate $O(\frac{1}{t})$ (Gower et al., 2019, theorem 3.2).

SGD vs. GD. Recall that n is the size of the dataset. Each iteration of gradient descent has a compute cost of $O(n)$, whereas SGD has a constant cost of $O(1)$. We see that even though each iteration of SGD is n times more efficient than an iteration of full batch gradient descent, its overall convergence rate is $O(\frac{1}{t})$, far worse than the linear rate of gradient descent $O(e^{-\frac{t}{\kappa}})$. Finding an algorithm with a cheap $O(1)$ iteration cost and a linear convergence rate seemed impossible until the advent of SAG (Roux et al., 2012) and variance reduction techniques.

2.2.2 Variance Reduction

Compared to the expected population risk in Eq. (2.20), the empirical risk Eq. (2.18) has a particular finite sum structure that SGD does not exploit. This fact is exploited – in the specific context of logistic regression – by the online exponentiated gradient (OEG) algorithm (Collins et al., 2008). OEG treats one sample at a time, so it has a constant iteration cost, independent of the dataset size. And yet, it enjoys a linear convergence rate.

OEG enjoyed a significant success on many problems, but it is not until the breakthrough work of Roux et al. (2012) that the true power of the finite sum structure was revealed. Roux et al. (2012) designed and analyzed a generic stochastic gradient-based algorithm with a cheap $O(1)$ iteration cost and a linear convergence rate. This algorithm is called stochastic averaged gradient or SAG. Similar to SGD, at each step it sample a datapoint x_i and computes its gradient $\nabla f(\theta_t, x_i)$. The difference is that it estimates the true gradient thanks to past gradients of each individual data points $\nabla f(\theta_{t_i}, x_i)$ where t_i is the last time that we sampled x_i. Finally the update of SAG writes

$$\theta_{t+1} = \theta_t - \frac{\gamma_t}{n} \sum_i \nabla f(\theta_{t_i}, x_i) \; . \tag{SAG}$$

Following this path, Defazio et al. (2014) introduced SAGA, a very similar algorithm with an unbiased gradient estimate allowing for more straightforward analysis. Concurrently Shalev-Shwartz and Zhang (2013b) analyzed SDCA, an algorithm maximizing a dual formulation for convex regularized problems (see §2.2.3). SDCA is a close cousin of OEG, and it enjoys the same constant iteration cost and linear convergence rate. In our first contribution, we improve upon SDCA and apply it to the challenging problem of conditional random fields (see §2.3.2).

Unfortunately, the memory footprint of SAG, SAGA or SDCA is $O(nd)$ in general (it can be reduced to $O(n)$ in many scenarios), which can quickly become prohibitive for large datasets or large models. Johnson and Zhang (2013) introduced stochastic variance reduced gradient (SVRG) to alleviate this issue. Instead of storing all past gradients, SVRG stores one past iterate θ_T along with its full batch gradient $\nabla F(\theta_T)$, and it applies the update

$$\theta_{t+1} = \theta_t - \gamma_t(\nabla f(\theta_t) - \nabla f(\theta_T) + \nabla F(\theta_T)) \; . \tag{SVRG}$$

Thus SVRG only needs $O(d)$ memory, but it needs twice more compute than plain SGD. As such, the variance reduction technique is most amenable to optimizing large models such as neural networks.

2.2.3 Fenchel Duality

As previously mentioned, and fully explained in our first contribution, SDCA operates on the dual formulation of

$$\min_{\theta} \frac{1}{n} \sum_{i} f(y_i, \theta^\top x_i) + \frac{\lambda}{2} \|\theta\|^2 \ . \tag{2.22}$$

However, what is this a dual formulation?

Convex Conjugates. To explain Fenchel duality properly, we need to introduce the convex conjugate of a function.

Definition 2.2.4 (convex conjugate)**.** *The convex conjugate of a function* $f : \mathbb{R}^d \to \mathbb{R} \cup +\infty$ *is defined by the pointwise formula*

$$f^*(y) := \max_x \langle y, x \rangle - f(x) \tag{2.23}$$

This transformation is a ubiquitous concept throughout Science. In thermodynamics and classical mechanics, it appears as the Legendre transform (a special case). In convex optimization and machine learning, we call it the Fenchel conjugate or the convex conjugate. At first sight, this definition seems arbitrary, but it admits geometrical interpretations along with many properties that make it a helpful tool. The author of this thesis produced several interactive tools to grasp a better understanding of convex conjugates: DualityViz and Dual Snakes.

One of the most interesting properties of convex conjugation is that for convex functions, the convex conjugate of the conjugate is equal to the function itself, i.e.,

$$f^{**} = f \ . \tag{2.24}$$

Fenchel Dual. Assume we want to solve a composite minimization problem

$$\min_x f(x) + g(\boldsymbol{A}x) \tag{2.25}$$

where $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ are convex functions and $\boldsymbol{A} : \mathcal{X} \to \mathcal{Y}$ is a linear operator. Under mild assumptions, this problem can be equivalently expressed with the convex conjugates of f and g

$$\begin{aligned} \min_x f(x) + g(\boldsymbol{A}x) &= \min_x \max_y f(x) + \langle Ax, y \rangle - g^*(y) && (2.26) \\ &\geq \max_y \min_x f(x) + \langle x, A^\top y \rangle - g^*(y) && (2.27) \\ &= \max_y -f^*(-A^\top y) - g^*(y) \ . && (2.28) \end{aligned}$$

This last line is known as the Fenchel dual of problem (2.25). We inverted min and max between the first and second line to reach it. Fenchel's duality theorem states sufficient conditions for this inequality to be an equality, in which case we say that strong duality holds. Fenchel duality is equivalent to Lagrange duality (Magnanti, 1974), but Fenchel's is more convenient for unconstrained problems or problems where the constraints are implicitly defined in the objective, whereas Lagrange's is more convenient for explicitly defined constraint.

SDCA and many other optimization algorithms directly store and update the dual variable y. It is well defined for generalized linear models, e.g., models defined with the exponential family.

2.3 Probabilistic Models

2.3.1 Exponential Families

Exponential families are among the simplest parametric models of distributions. To define an exponential family, take a variable x in $\mathcal{X}$ equipped with the base measure ν. Then extract a sufficient statistic $T(x) \in \mathbb{R}^d$. Then take the inner product between some parameter θ and $T(x)$. This inner product may be negative, so to ensure it is positive, take its exponential $e^{\langle \mathbb{E}[T(x)], \theta \rangle}$. The mass (for discrete random variables) or the density (for continuous random variables) with respect to ν is then defined to be proportional to this exponential

$$\boldsymbol{p}_\theta(x) \propto e^{\langle \mathbb{E}[T(x)], \theta \rangle} \nu(x) \ . \tag{2.29}$$

The logarithm of the normalization constant is known as the log-partition function

$$A(\theta) := \log \int e^{\langle \theta, T(x) \rangle} \nu(dx) \ . \tag{2.30}$$

The equation for the negative log-likelihood finally reads

$$f(\theta) := \mathbb{E}[-\log p_\theta(X)] = A(\theta) - \langle \mathbb{E}[T(X)], \theta \rangle \ . \tag{2.31}$$

Remark that f is convex, and it can be seen as a linear modification of the log-partition function A, which contains all the complexity. We provide more properties of these families in our third contribution.

We can always define $z = T(x)$, with μ the proper push forward modification of ν, in which case we say that Z belongs to the *natural exponential family* on μ (Morris, 1982).

The most common parametric distributions are exponential families: Categorical, Gaussians, Gamma, Wishart, Dirichlet, *etc.* A remarkable exception is the non-central Laplace $p(x) \propto e^{|x-\mu|}$ which cannot be expressed in this form.

GLM. Generalized linear models (GLM) are a powerful tool in supervised learning. Taking features x and labels y, a GLM models the conditional distribution $\boldsymbol{p}(y \mid x)$ with an exponential family whose parameter is a linear function of x. For simplicity, we consider the natural exponential family $T(y) = y$. The model writes

$$\boldsymbol{p}_\theta(y \mid x) = \exp(y^\top \boldsymbol{\theta} x - A(\boldsymbol{\theta} x)) \tag{2.32}$$

where $\boldsymbol{\theta}$ is a matrix of size $\dim(y) \times \dim(x)$.

In our first contribution, we study an algorithm for training a GLM for categorical distributions, e.g., logistic regression, with the number of categories growing exponentially with the input size. For this purpose, we use independence assumptions that are formalized by probabilistic graphical models.

2.3.2 Probabilistic Graphical Models

One of the most useful properties we can model about the natural distribution $\boldsymbol{p}(x)$ is the notion of (conditional) independence between variables. For instance, in a simple video game, two stacks of frames are often independent, given the stack of frames in between them. This kind of independence statements can be specified with graphs thanks to probabilistic graphical models – see Pearl (1988) for an historical reference, or Wainwright and Jordan (2008) or Koller and Friedman (2009) for a more recent review.

We start by presenting undirected graphical models. GLMs associated with undirected graphical models are known as the conditional random field (CRF), and they are at the core of our first contribution. Then we introduce directed graphical models, which are necessary to understand structural causal models, and our second contribution.

Undirected Graphical Models, a.k.a. Markov Random Fields

Let $\mathcal{G}$ be an undirected graph defined by its vertices $\mathcal{V} = \{1, \dots, d\}$ and its edges $(i, j) \in \mathcal{E}$.

Definition 2.3.1 (clique)**.** *The set $C = \{v_1, \dots, v_k\}$ is said to be a clique of $\mathcal{G}$ if and only if it forms a complete graph, e.g. $\forall i \neq j, (v_i, v_j) \in \mathcal{E}$.*

Definition 2.3.2 (maximal clique)**.** *A clique C is maximal if it is not contained in any clique, e.g. $\forall C', (C \subset C' \implies C'$ is not a clique$)$.*

We name $\mathcal{C}$ the set of maximal cliques of $\mathcal{G}$. We are now ready to define the independence statement.

Definition 2.3.3. *A distribution $\boldsymbol{p}$ is said to factor along $\mathcal{G}$ if and only if its density verifies*

$$\boldsymbol{p}(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{2.33}$$

where x_C is a vector containing the rows of x indexed by $i \in C$, and $\psi_C : \mathbb{R}^{|C|} \to \mathbb{R}$ are real-valued functions of $|C|$ elements. We refer to ψ_C as the potential of the clique C.

Exponential Graphical Model. If $\boldsymbol{p}_\theta$ belongs to the exponential family, then a sufficient condition for $\boldsymbol{p}_\theta$ to factor along $\mathcal{G}$ is for its sufficient statistic to decompose along with the cliques of $\mathcal{G}$, e.g.

$$T(x) = \sum_{C \in \mathcal{C}} T_C(x_C) \tag{2.34}$$

$$\implies \boldsymbol{p}_\theta(x) \propto \prod_{C \in \mathcal{C}} e^{\langle T_C(x_C), \theta \rangle} \ . \tag{2.35}$$

We exploit this fact in our first contribution.

Directed Graphical Models, a.k.a. Bayesian Networks

Directed probabilistic graphical models are also known as Bayesian Networks since Pearl (1985) coined this term. They are perhaps simpler to understand than undirected graphical models, but they are not easier to deal with.

Suppose we observe a random variable $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ with probability law $\boldsymbol{p}(X)$. We are also given a Directed Acyclic Graph (DAG) $\mathcal{G}$ with vertex $\mathcal{V} = \{1, \ldots, d\}$ and edges $\mathcal{E}$. We denote $\mathrm{Pa}(i)$ the parents of node i. This is the empty set if i has no parents.

Definition 2.3.4. *We say that $\boldsymbol{p}$ factorizes along $\mathcal{G}$ iff*

$$\boldsymbol{p}(X) = \prod_{i=1}^{d} \boldsymbol{p}(X_i | X_{\mathrm{Pa}(i)}) \ . \tag{2.36}$$

In other words, the only conditional dependencies of $\boldsymbol{p}$ are indicated by the edges of the graph G. The fewer edges in G, the more we know about X. In fact, if we know nothing about $\boldsymbol{p}$, we still know that we can write it as

$$\boldsymbol{p}(X) = \prod_{i=1}^{d} \boldsymbol{p}(X_i | X_{<i}) \tag{2.37}$$

by definition of conditional probability – modulo some positivity constraints. Consequently, a useful graph should have only a few edges, or equivalently a low degree.

Structure Learning. In unsupervised learning, either we posit that the data factorizes along with a graph and exploit this information to learn a density model $\boldsymbol{p}_\theta$ with fewer data. Either we set the goal of discovering these conditional independence structures. This goal is known as structure learning. Current solutions to this problem fall into two categories

1. Explicitly find out conditional independences with statistical testing and build the graph from there.

2. Use a scoring function to explore all possible graphs and keep the one with the highest score. The scoring function is often designed as the posterior probability of the structure given the data.

Directed graphical models have proven helpful in many modeling areas. However, they alone cannot predict what will happen if one of the variables is affected by some external stimuli. That is the topic of causal inference.

2.3.3 Causal Inference

Causal inference use directed graphical models to predict the effect of interventions in the world. We will now introduce two key elements of this theory: do-calculus and structural causal models.

Do-calculus

Assume we have data for kidney stone treatments performed in one hospital. For each patient, we know the treatment they received X, the outcome Y – did they successfully heal? – and the size of the stones they found during the surgery Z. A new patient arrives. We have to recommend the treatment that will maximize their chance of recovery. How should we process the data to make this decision?

This classic story is an instance of Simpson's paradox. The straightforward solution would be to recommend the treatment with the highest success rate in this example. However, it so happens that the treatment received by past patients was picked based on their symptoms, which were themselves a function of the stone sizes. In this example, the stone size is a *confounder* that affects both the treatment and the outcome. First, one should partition based on Z the data before aggregating the success rates. But why is that, and how to formalize that? The answer lies in the work of Judea Pearl (Pearl, 2009) and other statisticians. It can be formalized with the help of graphical models such as Figure 2.1.

The question we asked is an *interventional question*: what will happen *if we assign* $X = x$*?* This action effectively removes the observed statistical dependency between X and Z. The outcome distribution of this action should not be computed as the simple conditional probability $P(Y|X = x)$, but as another quantity that we will denote $P(Y|\,\mathrm{do}(x))$. The gold standard to estimate this quantity would

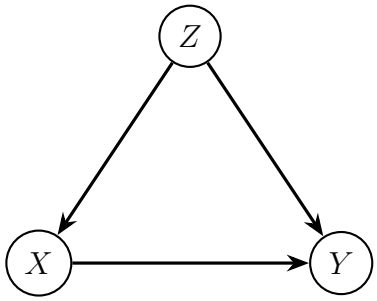


Figure 2.1 – The graph of causal relationships between treatments X, outcome Y and stone size Z. Z is a cause of both X and Y, which makes it a confounder.

be to perform a *randomized control trial*, where we blindly and randomly assign treatments to incoming patients, then observe and report success rate. However, we want to exploit the observed data to estimate this quantity. That is where the *do-calculus* comes into play. It is a set of rules based on graphs that transform do-statements such as $P(Y|\,\mathrm{do}(x))$ into an equation written in terms of observed probabilities. In the kidney stone example, we can estimate $P(Y|\,\mathrm{do}(x))$ from observational data thanks to the *backdoor adjustment formula*

$$P(Y|\,\mathrm{do}(x)) = \sum_z P(Y|x,z)P(z) \neq \sum_z P(Y|x,z)P(z|x) = P(Y|x)\ . \tag{2.38}$$

What is critical here is that Z is a cause of X. If instead X caused Z then causal effect and conditional would be equal $P(Y|\,\mathrm{do}(x)) = P(Y|x)$. Yet from a Bayesian network perspective, both arrow directions make a complete graph, which encodes the same absence of conditional independence. In other words, a causal graphical model encodes strictly more information than a Bayesian network.

The backdoor adjustment formula is the most famous instance of do-calculus, but more complex rules exist for complex graphs with both observed and unobserved variables. Quite recently, Huang and Valtorta (2012) proved that these rules are complete, meaning that if a do-statement can be expressed in terms of observed probabilities, then one will be able to find the right formula by applying these rules.

Structural Causal Models

Thanks to the rules of do-calculus, knowing the causal graph can be handy. So far, we have talked about this in a non-parametric setting, assuming we have direct access to the observed conditional probabilities $P(Y|X,Z)$. In reality, we need to parametrize these mechanisms. This is what a Structural Causal Model (SCM) is for. It describes a causal model by a set of unobserved independent exogenous noise variables $U_1,\ldots,U_d$, and a set of functions $f_1,\ldots,f_d$ such that

$$X_i = f_i(X_{\mathrm{Pa}(i)}, U_i), \forall i\ . \tag{2.39}$$

Given a DAG $\mathcal{G}$, these functions, and distributions for the exogenous noise, one can sample a vector X by sampling the noises and applying these functions in a topological order of $\mathcal{G}$.

Among other things, SCMs are helpful to answer *counterfactual questions*: what would have happened if I had given the other treatment to this patient? Counterfactuals are a major topic in the causality community, but they are not relevant to this thesis, so we will not cover this theory.

While the formalism of (2.39) may seem trivial at first, it becomes useful when one starts thinking about causal structure discovery. If we assume a parametric form for the functions, then the graphical structure can become identifiable, meaning that only one graph could have generated the observed data. One such example is if we assume f_i are linear and noises are non-Gaussian. However, the interest of these identifiability results is limited because, in general, we have no guarantee on the shape of the function that generated the data.

The SCM formalism enables us to think about a much deeper hypothesis: **Independent Causal Mechanisms**. This hypothesis postulates that knowing something about one mechanism does not provide any information about the others. That can be formalized by various means. One of them is Algorithmic Information Theory: the Kolmogorov Complexity of the set $\{f_1, \ldots, f_d\}$ [2] is on the same order of magnitude as the sum of the Kolmogorov Complexity of each function taken independently. Using this independence insight, one can devise algorithms that aim to find the data's causal structure. See Peters et al. (2017) for a book on this topic.

Our second contribution addresses an idea to discover causal structure from interventional data, e.g., data coming from (possibly unknown) interventions. We have now provided all the key elements to understand this thesis. Let us emphasize that causal inference and causal discovery are taking more and more space in machine learning. We refer the reader to (Schölkopf, 2019) for a modern review of the literature.

[2] We do not include the exogenous noise distributions for simplicity.

3 Adaptive Stochastic Dual Coordinate Ascent for Conditional Random Fields

Prologue to the First Contribution

Article Details

Adaptive Stochastic Dual Coordinate Ascent for Conditional Random Fields. *Rémi Le Priol, Alexandre Piché and Simon Lacoste-Julien.* Published at UAI 2018 (Le Priol et al., 2018).

Contributions of the Authors

Rémi Le Priol wrote most of the code (and most of the bugs) and ran most experiments. He also found the proof for acceleration under adaptive sampling. Alexandre Piché contributed to the code and ran experiments. Simon Lacoste-Julien provided supervision. All authors contributed to the writing of the paper.

Abstract

This work investigates the training of conditional random fields (CRFs) via the stochastic dual coordinate ascent (SDCA) algorithm of Shalev-Shwartz and Zhang (2016). SDCA enjoys a linear convergence rate and a strong empirical performance for binary classification problems. However, it has never been used to train CRFs. Yet it benefits from an "exact" line search with a single marginalization oracle call, unlike previous approaches. In this paper, we adapt SDCA to train CRFs and enhance it with an adaptive non-uniform sampling strategy based on block duality gaps. We perform experiments on four standard sequence prediction tasks. SDCA demonstrates performances on par with state of the art and improves over it on three of the four datasets, which have in common the use of sparse features.

3.1 Introduction

The conditional random field (CRF) model (Lafferty et al., 2001) is a common tool in natural language processing and computer vision for structured prediction.

The optimization of this model is notoriously challenging. Schmidt et al. (2015) describes a practical implementation of the stochastic average gradient (SAG) algorithm (Roux et al., 2012) for CRFs and proposes a non-uniform sampling scheme that boosts performance. This algorithm (SAG-NUS) is currently state of the art for CRFs optimization, and we refer to Schmidt et al. (2015) for a detailed review of competing methods.

Deterministic (batch) methods such as L-BFGS (Sha and Pereira, 2003; Wallach, 2002) have a linear convergence rate, but the cost per iteration is high. On the other hand, the online exponentiated gradient method (OEG) (Collins et al., 2008) and SAG are both members of a family of algorithms with cheap stochastic updates and linear convergence rates, and they have both been applied to the training of CRFs. They are called variance-reduced algorithms because their common point is to use memory to reduce the variance of the stochastic update direction as they get closer to the optimum. Johnson and Zhang (2013) coined the name stochastic variance reduced gradient (SVRG), and Defazio et al. (2014) unified the family.

The stochastic dual coordinate ascent (SDCA) algorithm proposed by Shalev-Shwartz and Zhang (2013b, 2016) is a member of this family that has not yet been applied to CRFs. It is closely related to OEG in that it also does block-coordinate ascent on the dual objective. Yet an interesting advantage of SDCA over OEG (and SAG) is that the form of its update makes it possible to perform an "exact" line search with only *one* call to the *marginalization oracle*, i.e., the computation of the marginal probabilities for the CRF. This contrasts with both SAG and OEG, where each step size change requires a new call to the marginalization oracle. We thus propose in this paper to investigate the performance of SDCA for training CRFs.

Contributions. We adapt the multi-class variant of SDCA to the CRF setting by considering the marginal probabilities over the cliques of the graphical model. We provide a novel interpretation of SDCA as a relaxed fixed point update and highlight the block separability of the duality gap. We propose to enhance SDCA with an adaptive non-uniform sampling strategy based on the block gaps and analyze its theoretical convergence improvement over uniform sampling. We compare the state-of-the-art methods on four prediction tasks with a sequence structure. SDCA with uniform sampling performs comparably with OEG and SAG. When SDCA is enhanced with the adaptive sampling strategy, it outperforms its competitors in terms of the number of parameters updates on three of the tasks. These three tasks are all about natural language with sparse handcrafted features. We hypothesize that the efficiency of the dual methods can be related to the sparsity of these features.

Related work. Our proposed gap sampling strategy is similar to the one from Osokin et al. (2016) in the context of SDCA applied to the structured SVM objective, which reduces to the block-coordinate Frank-Wolfe (BCFW) algorithm (Lacoste-Julien et al., 2013). Dünner et al. (2017) recently analyzed a general adaptive sampling scheme for approximate block coordinate ascent that

generalizes SDCA. Their proposed sampling scheme (which chooses the biggest gap) was motivated in the different contexts of mixed GPU and CPU computations, which does not apply to our setting. Our proposed practical strategy considers the gaps' staleness and is more robust in our experimental setting. Csiba et al. (2015) proposes an adaptive sampling scheme for SDCA for binary classification, which unfortunately cannot be generalized to the CRF setting due to an intractable computation. Closely related to our work is Perekrestenko et al. (2017), who analyzed several adaptive sampling strategies for a generalization of the primal-dual SDCA setup, including our proposed gap sampling scheme. However, their analysis was focused on the single coordinate descent method (e.g., binary SDCA) and on sublinear convergence results obtained when strong convexity is not assumed. Instead, we cover the block-coordinate approach relevant to CRFs, and one of our notable results is to show that the linear convergence rate for gap sampling **dominates** the one for uniform sampling, in contrast to what happens in the sublinear regime studied by Perekrestenko et al. (2017).

Outline. We review the optimization problem for CRFs as well as provide novel insights on the primal-dual optimization structure in Section 3.2. We present SDCA for CRFs in Section 3.3 and discuss important implementation aspects in Section 3.4. We present and analyze various adaptive sampling schemes for SDCA in Section 3.5. We provide experiments in Section 3.6 and discuss the implications in Section 3.7.

3.2 Conditional Random Fields

This section reviews the CRF model and its associated primal and dual optimization problems. We then derive some interesting properties which motivate several optimization algorithms.

3.2.1 Definition

A CRF models the conditional probability of a structured output $y \in \mathcal{Y}$ (e.g. a sequence) given an input $x \in \mathcal{X}$ with a Markov random field that uses an exponential family parameterization with sufficient statistics $F(x, y) \in \mathbb{R}^d$ and parameters $\boldsymbol{w} \in \mathbb{R}^d : p(y|x; \boldsymbol{w}) \propto \exp(\boldsymbol{w}^\top F(x, y))$. The feature vector F decomposes as a sum over the cliques $C \in \mathcal{C}$ of the graphical model for y: $F(x, y) = \sum_C F_C(x, y_C)$, where y_C denotes the subset of coordinates of y selected by the indices from the set C. See Figure 3.1 for an illustration.

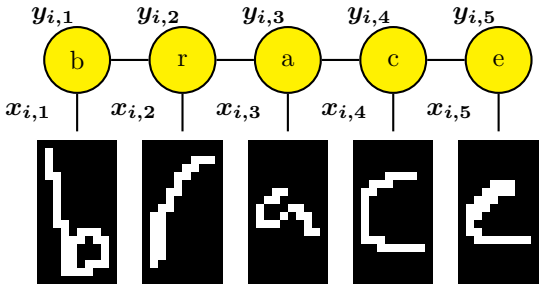


Figure 3.1 – Example of graphical model for the optical character recognition (OCR) task. We want to exploit the structure of the word to predict that $y_{i,5}$ is an "e" and not a "c". This can be done by working on the pairs $y_{i,\{t,t+1\}} = (y_{i,t}, y_{i,t+1})$, the cliques of that model.

3.2.2 Primal Problem

We have a data set $(x_i, y_i)_{i\in[1,n]}$ of n i.i.d. input and structured output pairs. The parameter is learned by minimizing the ℓ_2-regularized negative log-likelihood:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} -\log\left(p(y_i|x_i;\boldsymbol{w})\right) . \tag{3.1}$$

We now rewrite it using the notation for the SDCA setup for multi-class classification from Shalev-Shwartz and Zhang (2016). Denote $M_i = |\mathcal{Y}_i|$ the number of labelings for sequence i. Denote A_i the $d \times M_i$ matrix whose columns are the *corrected features* $\{\psi_i(y) := F(x_i, y_i) - F(x_i, y)\}_{y\in\mathcal{Y}_i}$. Denote also $\phi_i(s) := \log\left(\sum_{y\in\mathcal{Y}_i}\exp(s_y)\right)$ the log-partition function for the scores $s \in \mathbb{R}^{M_i}$. The negative log-likelihood can be written $-\log(p(y_i|x_i;\boldsymbol{w})) = \phi_i(-A_i^\top \boldsymbol{w})$. The primal objective function to minimize over $\boldsymbol{w} \in \mathbb{R}^d$ thus becomes:

$$\mathcal{P}(\boldsymbol{w}) := \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\phi_i(-A_i^\top\boldsymbol{w}) . \tag{3.2}$$

3.2.3 Dual Formulation

The above minimization problem (3.2) has an equivalent *Fenchel convex dual* problem (Lebanon and Lafferty, 2002). Denote Δ_M the probability simplex over M elements. Denote $\alpha_i \in \Delta_{M_i}$ the set of dual variables for a given x_i. The dual problem directly handles the probability of the labels for the training set. The dual objective to maximize over the choice of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \Delta_{|\mathcal{Y}_1|} \times \ldots \times \Delta_{|\mathcal{Y}_n|}$ is:

$$\mathcal{D}(\boldsymbol{\alpha}) := -\frac{\lambda}{2}\|\frac{1}{n\lambda}\sum_i A_i\alpha_i\|^2 + \frac{1}{n}\sum_{i=1}^{n} H(\alpha_i) , \tag{3.3}$$

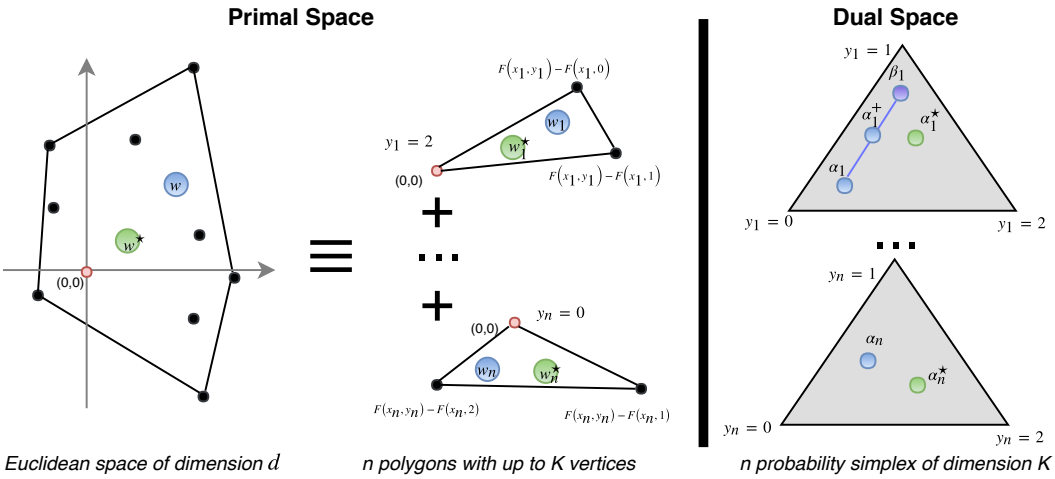


Figure 3.2 – **Left:** primal parameters $\boldsymbol{w}$ and $\boldsymbol{w}^*$ are a convex combination of corrected features ψ_i. In fact they are the average of n barycenters os smaller polygons with K vertices as per (3.4). **Right:** dual parameters live in n K-simplex. SDCA updates one simplex at a time with a relaxed fixed point iteration (3.16).

where $H(\alpha_i) := -\sum_{y \in \mathcal{Y}_i} \alpha_i(y) \log(\alpha_i(y))$ is the entropy of the probability distribution α_i. The negative entropy appears as the convex conjugate of the softmax: $-H = \phi^*$. An illustration of primal and dual parameters is provided in Figure 3.2

3.2.4 Optimality Conditions

We define the *conjugate weight* function $\hat{w}$ as follows:

$$\hat{w}(\boldsymbol{\alpha}) := \frac{1}{n\lambda} \sum_i A_i \alpha_i = \frac{1}{\lambda n} \sum_{i=1}^n \mathbb{E}_{y \sim \alpha_i}[\psi_i(y)] \tag{3.4}$$

$$= \frac{1}{\lambda} \left(\frac{1}{n} \sum_{i=1}^n F(x_i, y_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{y \sim \alpha_i}[F(x_i, y)] \right) . \tag{3.5}$$

It is the difference between the average of the ground truth features, and the average of the expected features for the dual variable, up to a factor $\frac{1}{\lambda}$. One can show that $\hat{w}(\boldsymbol{\alpha}^\star) = \boldsymbol{w}^\star$ where $\boldsymbol{w}^\star$ and $\boldsymbol{\alpha}^\star$ are respectively the optimal primal parameters and the optimal dual parameters.

We also define *conjugate* probabilities $\hat{\alpha}_i$ as follows:

$$\forall i, \quad \hat{\alpha}_i(\boldsymbol{w}) := \nabla_s \phi_i(-A_i^\top \boldsymbol{w}) = p(.|x_i; \boldsymbol{w}). \tag{3.6}$$

We get another optimality condition $\hat{\alpha}(\boldsymbol{w}^\star) = \boldsymbol{\alpha}^\star$. These two optimality conditions can be deduced directly from the structure of the duality gaps.

3.2.5 Duality Gaps

Note that $\mathcal{P}(\boldsymbol{w}) \geq \mathcal{D}(\boldsymbol{\alpha})$ is always true, with equality at the optimum. The *duality gap* is defined by:

$$g(\boldsymbol{w}, \boldsymbol{\alpha}) = \mathcal{P}(\boldsymbol{w}) - \mathcal{D}(\boldsymbol{\alpha})\,. \tag{3.7}$$

Note that we can rewrite the primal gradient as following:

$$\nabla \mathcal{P}(\boldsymbol{w}) = \lambda(\boldsymbol{w} - \hat{w} \circ \hat{\alpha}(\boldsymbol{w}))\,. \tag{3.8}$$

One can verify that:

$$\begin{aligned} g(\boldsymbol{w}, \hat{\alpha}(\boldsymbol{w})) &= \frac{\lambda}{2}\|\boldsymbol{w} - \hat{w}(\hat{\alpha}(\boldsymbol{w}))\|^2 && (3.9)\\ &= \frac{1}{2\lambda}\|\nabla \mathcal{P}(\boldsymbol{w})\|^2\,. && (3.10) \end{aligned}$$

This structure of the gap for the primal weights and its dual conjugate probabilities have an equivalent in the dual. Denote the Fenchel duality gap of ϕ_i for the scores $s_i = -A_i^T \boldsymbol{w}$ and probabilities $\boldsymbol{\alpha}_i$:

$$F_i(s_i, \alpha_i) := \phi_i(s_i) + \phi_i^*(\alpha_i) + s_i^T \alpha_i \geq 0. \tag{3.11}$$

The positivity comes from the definition of convex conjugates. The gap is zero when s_i and α_i are conjugate variables for ϕ_i, e.g. $\alpha_i = \nabla \phi_i(s_i)$. For any smooth loss ϕ_i, the duality gap between $\hat{w}(\boldsymbol{\alpha})$ and $\boldsymbol{\alpha}$ decomposes as a sum of Fenchel gaps (Shalev-Shwartz and Zhang, 2013a):

$$g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \frac{1}{n}\sum_i F(-A_i^T \hat{w}(\boldsymbol{\alpha}), \alpha_i). \tag{3.12}$$

The log-sum-exp and the entropy are a special pair of conjugates. Their Fenchel duality gap is also equal to the Bregman divergence generated by $\phi_i^* = -H$, the Kullback-Leibler divergence: $F_i(s_i, \alpha_i) = D_{KL}(\alpha_i || \nabla \phi_i(s_i))$. Writing this for the same pair of conjugate variables yields:

$$g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \frac{1}{n}\sum_i D_{KL}(\alpha_i || \hat{\alpha}_i(\hat{w}(\boldsymbol{\alpha})). \tag{3.13}$$

The duality gaps (3.9) and (3.13) are typically used to monitor the optimization. In Appendix 3.D, we explain how one can transfer a convergence guarantee on the primal or dual suboptimality to a convergence guarantee on the duality gap.[1] Moreover, the block-separability of gaps from (3.13) can motivate an adaptive sampling scheme, as we describe in Section 3.5.

[1] This implies that convergence results on the dual problem directly translates to convergence results on the primal and vice-versa; a fact apparently missed in the linear rate comparison of Schmidt et al. (2015).

3.2.6 Interpretation

The primal formulation chooses a $\boldsymbol{w}$ with a small norm to maximize the conditional probability of observing the labels. Conversely, the dual formulation chooses conditional probabilities of the labels so as to minimize the ℓ_2 distance between the expected features and empirical expectation of the ground truth features. The optimal distribution would be the empirical distribution, if not for the entropic regularization that favors more uniform probabilities. This is the regularized version of the classical duality between maximum-likelihood and maximum-entropy for exponential families.

The optimality conditions show that the solution of the primal Problem (3.2) is also a *fixed point* for the function $\hat{w} \circ \hat{\alpha}$. Because of the gradient form (3.8), the gradient descent update can also be written as a *relaxed* fixed point update:

$$\boldsymbol{w}^+ = \boldsymbol{w} - \gamma \nabla \mathcal{P}(\boldsymbol{w}) \tag{3.14}$$

$$= (1 - \gamma\lambda)\boldsymbol{w} + \gamma\lambda \; \hat{w} \circ \hat{\alpha}(\boldsymbol{w}) \, . \tag{3.15}$$

The algorithm SDCA described in the next section also admits a relaxed fixed point update on the block α_i (see (3.16)). More generally, optimization algorithms for Problem (3.2) can often be interpreted as a back and forth between the conjugate variables w and $\hat{w}(\hat{\alpha}(\boldsymbol{w}))$ (primal methods) or α and $\hat{\alpha}(\hat{w}(\boldsymbol{\alpha}))$ (dual methods). For instance, one could interpret OEG as a relaxed fixed point iteration over the score variables $s_i = -A_i^T \boldsymbol{w}$.

$$\begin{array}{ccc} \boldsymbol{w} & \xrightarrow{\hat{\alpha}} & \left(\nabla_s \phi_i(-A_i^T \boldsymbol{w})\right)_{i=1}^n \\ \uparrow & & \downarrow \\ \frac{1}{n\lambda} \sum_i A_i \alpha_i & \xleftarrow{\hat{w}} & \boldsymbol{\alpha} \end{array}$$

Most of the results presented in this section and in Section 3.5 can be transposed to other kinds of loss and regularization, under some regularity assumptions. Our focus in this paper is the application of SDCA to CRF models, and thus we focused the discussion on the log-likelihood setting and the ℓ_2 norm, which are widely used.

3.3 Stochastic Dual Coordinate Ascent

We first describe the SDCA in its general setting and then describe the necessary modifications for training a CRF.

Algorithm 1 Prox-SDCA (option II) called SDCA here

Initialize $\alpha_i^{(0)} \in \Delta_{M_i}, \forall i$
Let $\boldsymbol{w}^{(0)} = \hat{w}(\boldsymbol{\alpha}^{(0)}) = \frac{1}{\lambda n} \sum_i A_i \alpha_i$
for $t = 0, 1 \ldots$ **do**
  Sample i uniformly at random in $\{1, \ldots, n\}$
  Let $\beta_i := \hat{\alpha}_i(\boldsymbol{w}) = \nabla_s \phi(-A_i^T \boldsymbol{w})$
  Let $\delta_i = \beta_i - \alpha_i^{(t)}$ {dual ascent direction}
  Let $\boldsymbol{v}_i = \frac{1}{\lambda n} A_i \delta_i$ {primal direction}
  Solve Equation (3.17) to get γ^* {Line Search}
  Update $\alpha_i^{(t+1)} := \alpha_i^{(t)} + \gamma^* \delta_i$
  Update $\boldsymbol{w}^{(t+1)} := \hat{w}(\boldsymbol{\alpha}^{(t+1)}) = \boldsymbol{w}^{(t)} + \gamma^* \boldsymbol{v}_i$

3.3.1 General Setting

The stochastic dual coordinate ascent algorithm (SDCA) updates one dual coordinate at a time so as to maximize the dual objective. SDCA was originally proposed for binary classification (Shalev-Shwartz and Zhang, 2013b) where each dual variable α_i lives in $\Delta_2 = [0, 1]$. In this case, it is possible to do exact coordinate maximization of the dual objective over a single α_i with standard one-dimensional optimization.

However, there is no simple way to maximize the dual objective over the block $\alpha_i \in \Delta_K$ in the multi-class setting. The algorithm with the surprising name of Proximal-SDCA[2], option II (Shalev-Shwartz and Zhang, 2016) proposes a solution to this problem. It updates α_i in a clever direction derived from the primal-dual relationship, which amounts to a relaxed fixed point update. See Figure 3.2 for an intuition and Algorithm 1 for the details.

We now describe the idea. At all time, we maintain the pair of dual and primal variables $(\boldsymbol{\alpha}, \boldsymbol{w} = \hat{w}(\boldsymbol{\alpha}))$. At each step, we sample a training point i. We compute $\beta_i = \nabla_s \phi_i(-A_i^T \boldsymbol{w}) = \hat{\alpha}_i \circ \hat{w}(\boldsymbol{\alpha})$, the next fixed point iterate. We then define the dual ascent direction by $\delta_i := \beta_i - \alpha_i$. Finally we update the block α_i with the right step size so as to increase the dual objective $\mathcal{D}(\boldsymbol{\alpha})$ using a relaxed fixed point update:

$$\alpha_i^+ \leftarrow \alpha_i + \gamma \delta_i = (1 - \gamma)\alpha_i + \gamma \hat{\alpha}_i \circ \hat{w}(\boldsymbol{\alpha}) \,. \tag{3.16}$$

The dual ascent direction is guaranteed to increase $\mathcal{D}(\boldsymbol{\alpha})$, unless $\delta_i = 0$ (this actually means that the block is already optimal, see (3.13)). The primal weights $\boldsymbol{w} = \hat{w}(\boldsymbol{\alpha})$ are related to $\boldsymbol{\alpha}$ by a linear transformation. Define the primal direction $\boldsymbol{v}_i = \frac{1}{\lambda n} A_i \delta_i \in \mathbb{R}^d$. One can update the weights directly: $\boldsymbol{w}^+ \leftarrow \boldsymbol{w} + \gamma \boldsymbol{v}_i$.

The step size $\gamma \in [0, 1]$ is either fixed or found via line search. In practice, the fixed step size for which convergence is guaranteed is really small. The line search

[2]We call it SDCA in the rest of this paper

is relatively cheap as we are looking at only one block:

$$\gamma^* := \underset{\gamma \in [0,1]}{\operatorname{argmax}} -\phi_i^*(\alpha_i + \gamma\delta_i) - \frac{\lambda n}{2}\|\boldsymbol{w} + \gamma \boldsymbol{v}_i\|^2. \tag{3.17}$$

Note that one can decompose the quadratic term and precompute $\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle$ and $\|\boldsymbol{v}_i\|^2$ to accelerate the optimization. The bottleneck remains the computation of ϕ_i^* (and its derivatives).

3.3.2 Adaptation to CRF

In the CRF setting, the dual variable α_i is exponentially large in the input size x_i. For a sequence x_i of length T where each node can take up to K values, the number of possible labels is $M_i = |\mathcal{Y}_i| = K^T$. It might not even fit in memory. Instead, the standard approach used in OEG and SAG is to consider the marginal probabilities $(\mu_C)_{C \in \mathcal{C}}$ on the cliques of the graphical model. Similarly, we replace $\boldsymbol{\alpha}$ by $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$, where $\mu_i \in \prod_C \Delta_C$ is the concatenation of all the clique marginal vectors for the sample i. For the same sequence x_i, this reduces the memory cost to $K^2(T-1)$ for the pair marginals. We denote $m_i = \sum_C |\mathcal{Y}_{i,C}|$ this new memory fingerprint. For a sequence long enough, we have $m_i \ll M_i$. The associated weight vector can still be expressed as function of $\boldsymbol{\mu}$ thanks to the separability of the features:

$$\hat{w}(\boldsymbol{\mu}) = \frac{1}{\lambda n} \sum_i \sum_C \mathbb{E}_{\mu_{i,C}}[\psi_{i,C}] = \frac{1}{\lambda n} \sum_i B_i \mu_i, \tag{3.18}$$

where $B_i = (\psi_{i,C}(y_C))_{C, y_C} \in \mathbb{R}^{d \times m_i}$ is the horizontal concatenation of the cliques feature vectors.

Now, assume that the graph has a *junction tree* structure $T = (\mathcal{C}, \mathcal{S})$ (Koller and Friedman, 2009, Def. 10.3), where $\mathcal{C}$ is the set of maximal cliques and $\mathcal{S}$ the set of separators. We can then run message passing on the junction tree to infer the new marginals given weights $\boldsymbol{w}$: $\hat{\mu}_i(\boldsymbol{w}) = p(y_C = .|x_i; \boldsymbol{w})$. We can also now recover the joint probability $\alpha_i(y)$ as a function of its marginals $\mu_{i,C}$ (Koller and Friedman, 2009, Def. 10.6):

$$\alpha_i(y) = \frac{\prod_{C \in \mathcal{C}} \mu_{i,C}(y_C)}{\prod_{S \in \mathcal{S}} \mu_{i,S}(y_S)}. \tag{3.19}$$

Equation (3.19) in turn allows us to compute the entropy and the divergences of the joints, using only the marginals. Let μ_i and ν_i be the marginals of respectively α_i and β_i, then the entropy and the Kullback-Leibler divergence are given by:

$$\tilde{H}(\mu_i) := H(\alpha_i) = \sum_C H(\mu_{i,C}) - \sum_S H(\mu_{i,S}) \tag{3.20}$$

Algorithm 2 SDCA for CRF

Initialize $\mu_i^{(0)} \in \prod_C \Delta_C$ consistently $\forall i$ {use (3.23)}
Set $\boldsymbol{w}^{(0)} := \hat{w}(\boldsymbol{\mu}^{(0)}) = \frac{1}{\lambda n} \sum_i B_i \mu_i^{(0)}$ {See (3.18)}
(Optional) Let $g_i = 100, \forall i$
for $t = 0, 1 \ldots$ **do**
  Sample i uniformly at random in $\{1, \ldots, n\}$
  (Alternatively) Sample i proportionally to g_i
  Let $\nu_{i,C}(y_C) := p(y_C | x_i; \boldsymbol{w}^{(t)}), \forall C \in \mathcal{C}$ {**oracle**}
  (Optional) Let $g_i = \tilde{D}(\mu_i || \nu_i)$ {duality gap (3.21)}
  Let $\delta_i = \nu_i - \mu_i^{(t)}$ {ascent direction}
  Let $\boldsymbol{v}_i = \frac{1}{\lambda n} \hat{w}(\delta_i)$ {primal direction}
  Solve Equation (3.22) to get γ^* {Line Search}
  Update $\mu_i^{(t+1)} := \mu_i^{(t)} + \gamma^* \delta_i$
  Update $\boldsymbol{w}^{(t+1)} := \hat{w}(\boldsymbol{\mu}^{(t+1)}) = \boldsymbol{w}^{(t)} + \gamma^* \boldsymbol{v}_i$

and

$$\tilde{D}(\mu_i || \nu_i) := D_{KL}(\alpha_i || \beta_i) = \sum_C D_{KL}(\mu_{i,C} || \nu_{i,C}) - \sum_S D_{KL}(\mu_{i,S} || \nu_{i,S}). \tag{3.21}$$

With this expression of the entropy (3.20), we can compute the dual objective, and thus perform the line search:

$$\gamma^* = \operatorname*{argmax}_{\gamma \in [0,1]} \tilde{H}(\mu_i^{(t)} + \gamma \delta_i) - \frac{\lambda n}{2} \|\boldsymbol{w}^{(t)} + \gamma \boldsymbol{v}_i\|^2. \tag{3.22}$$

With the Kullback-Leibler divergence (3.21), we can efficiently compute the individual duality gaps from (3.13). Algorithm 2 describes this variation of SDCA, with as an option a non-uniform sampling strategy defined in Section 3.5.3.

3.4 Implementation

We provide in Appendix 3.A a discussion of various important implementation aspects summarized here.

1. The initialization of dual methods for CRFs can significantly influence their performance. As explained in Appendix 3.A, we use:

$$\boldsymbol{\alpha}^{(0)} := \varepsilon \boldsymbol{u} + (1 - \varepsilon) \boldsymbol{\delta}\,, \tag{3.23}$$

   where $\boldsymbol{u}$ is the uniform distribution on each block, $\boldsymbol{\delta}$ is a unit mass on each ground truth label, and ε is a small number.

2. Storing the dual variable may be expensive, and one should allocate a decent amount of memory.
3. The line search requires computing the entropy of the marginals. This is costly, and we used the Newton-Raphson algorithm to minimize the number of iterations. This in turn requires storing the logarithm of the dual variable.

3.5 Adaptive Sampling for SDCA

Recently, there has been a lot of attention on non-uniform sampling for stochastic methods. The general goal is to sample more frequently points that are harder to classify and can bring more progress on the objective. These methods are said to be *adaptive* when the sampling probability changes during the optimization. SDCA itself has had several adaptive schemes proposed. In the following, we attempt to explain and relate these methods and suggest new schemes that work well on our problem.

3.5.1 Ascent Lemma

We start by restating the ascent lemma from Equation (25) in Shalev-Shwartz and Zhang (2013a). This lemma inspires and supports all the strategies.

Ascent after sampling i: At iteration t, if we sample i and take a step of size $\gamma_i \in [0, 1]$, we can lower bound the resulting dual improvement:

$$\begin{aligned} & n(\mathcal{D}(\boldsymbol{\alpha}^+) - \mathcal{D}(\boldsymbol{\alpha})) \\ & \geq \gamma_i \underbrace{\left[\phi(-A_i^T \boldsymbol{w}) + \phi^*(\alpha_i) + \boldsymbol{w}^T A_i \alpha_i\right]}_{\text{Fenchel gap}=:g_i} + \gamma_i \left(\frac{(1-\gamma_i)}{2} - \frac{\gamma_i R_i}{2\lambda n} \right) \|\beta_i - \alpha_i\|_1^2 \end{aligned} \tag{3.24}$$

where $R_i := \|A_i\|_{1\to 2}^2 = \max_{y \in \mathcal{Y}_i} \|\psi_i(y)\|_2^2$ is the squared radius of the corrected features for sample i.

Note that compared to the original text, we used the fact that the regularizer is the ℓ_2 norm and the loss is 1-smooth with respect to the ℓ_∞ norm. We define $R := \max_i R_i$, $\bar{R} := \frac{1}{n}\sum_i R_i$ and $\bar{g} := \frac{1}{n}\sum_i g_i$ the true duality gap (see (3.11)-(3.12)). We also introduce $L_i := \lambda + \frac{R_i}{n}$ an upper bound on the smoothness of loss i plus regularizer for the ℓ_2 norm. We recall from Section 3.2.5 that $g_i = D_{KL}(\alpha_i || \beta_i)$ (3.13). We give the name *residual* to $d_i := \|\beta_i - \alpha_i\|_1^2$.

This lemma is derived with standard assumptions and inequalities on the smoothness of the loss and the strong convexity of the regularizer. The first term of the lower bound is the ascent guarantee, while the other term gives a condition on the

step size to ensure progress. We refer the reader to the original paper for more details.

To get the expected progress (conditioned on the past) after sampling with probability $\boldsymbol{p}$, we simply need to take the sum of the inequality above after multiplying both sides by p_i. Our goal is to maximize this lower bound by choosing the right probability $\boldsymbol{p}$ and step size $\boldsymbol{\gamma}$. To be able to conclude the proof with the original method, we also want some constants time the duality gap $\bar{g}$ to appear in the lower bound – the gap is lower bounded by the dual suboptimality, and thus this constant will give the linear rate of convergence. The lemma can then transpose this result from the dual sub-optimality to the duality gap as described in Appendix 3.D. From there on, there are two general approaches: importance sampling and duality gap sampling.

3.5.2 Importance and Residual Sampling

With the importance sampling approach, the goal is to set the step size and the probability so that they cancel each other out: $\gamma_i = \frac{\gamma}{p_i}$. One then gets an unbiased estimate of the true duality gap from (3.13) as the first term of the upper bound. What is left is maximizing the second term with respect to $\boldsymbol{p}$. This is the approach proposed by Zhao and Zhang (2015) (Importance Sampling, left term below) and generalized by Csiba et al. (2015) (Residual sampling, a.k.a. AdaSDCA for binary classification, right term):

$$p_i \propto L_i \quad \text{or} \quad p_i \propto d_i \sqrt{L_i}. \tag{3.25}$$

These sampling schemes somehow allow to maximize the second term of (3.24). Intuitively, they replace a dependency on R in the convergence rate by a dependency on $\bar{R}$. They can give good results on binary and multi-class logistic regression. There are a few issues though.

- One needs an accurate estimate of the L_i.
- Importance sampling is not adaptive.
- In the CRF setting, the residual is $d_i = \|\beta_i - \alpha_i\|_1^2$. It is the squared ℓ^1 norm of a vector of exponential size. We are not aware of any trick to compute it efficiently.

3.5.3 Gap Sampling

To make sure that the second term is positive, the original proof of uniform SDCA sets $\gamma_i = \gamma = (1 + \frac{R}{\lambda n})^{-1}$ to obtain:

$$n\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \gamma \sum_i p_i g_i. \tag{3.26}$$

Assuming a full knowledge of the duality gaps g_i, the optimal decision is sampling the point with the maximum duality gap. This was done by Dünner et al. (2017) in the context of multi-class classification on a pair CPU-GPU. While the GPU computes the update, the CPU updates as many duality gaps as possible. This leads to impressive acceleration over massive datasets.

However, this is not our current setting. We know and update only one gap at a time (for efficiency). Because of the staleness of the gaps, our experiments with this method did not even converge for the most part (see Section 3.6.3). We need a more robust method.

We take inspiration from what was done by Osokin et al. (2016) to improve the Block-Coordinate Frank-Wolfe (BCFW) algorithm (Lacoste-Julien et al., 2013). We propose to bias sampling towards examples whose duality gaps are large: $p_i \propto g_i$. If we know all the duality gaps, the expected improvement reads:

$$n\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \chi(\boldsymbol{g})^2 \, \gamma \, \bar{g}, \tag{3.27}$$

where $\chi(\boldsymbol{g}) = \sqrt{\frac{\frac{1}{n}\sum_i g_i^2}{\bar{g}^2}} \in [1, \sqrt{n}]$ is the non-uniformity of the duality gaps, as defined in Osokin et al. (2016, Section 3.1). The value $\chi(\boldsymbol{g})^2\gamma$ is the value that will appear in the linear convergence rate of this method. It means that the convergence rate for gap sampling **dominates** the one for uniform sampling. This is different from what was observed for BCFW, where they could not prove dominance in general.

In practice, we use stale estimates of the gaps, and there are no convergence guarantees. We discuss this issue more in section 3.6.3.

We also explored a combination of gap sampling and importance sampling. We could get a similar convergence rate where a trade-off appeared between the mean smoothness and the non-uniformity. We detail these considerations as a technical report in Appendix 3.F for the interested reader.

3.6 Experiments

We conducted these experiments to answer three questions: (1) How does the line search influence SDCA? (2) How do the non-uniform sampling schemes compare with each other? and (3) How does SDCA compare with SAG and OEG on sequence prediction?

3.6.1 Experimental Setting

We applied the experimental setup outlined by Schmidt et al. (2015). We implemented SDCA to train a classifier on four CRF training tasks: (1) the optical character recognition (OCR) dataset (Taskar et al., 2004), (2) the CoNLL-2000

Table 3.1 – Dataset summary. d is the dimension of $\boldsymbol{w}$. n is the number of data points (sequences). N is the number of nodes (e.g. sum of sequences length). K is the number of possible labels for each node. A is the number of attributes (see Appendix 3.B). a is the maximum number of attributes extracted from one node. Mem. is the memory required by the pairwise marginals stored as float 64. The pairwise marginals dominate the memory cost.

Dataset	OCR	CONLL	NER	POS
d	4,082	1.6×10^6	2.8×10^6	8.6×10^6
n	6,202	8,936	15,806	38,219
N	52,827	2.1×10^5	2×10^5	9.1×10^5
K	26	22	9	45
A	128	74,658	3.1×10^5	1.9×10^5
a	128	19	20	13
Mem.(GiB)	0.2	0.7	0.1	13

shallow parse chunking dataset (CONLL), (3) the CoNLL-2002 Dutch named-entity recognition dataset (NER), and (4) a part-of-speech (POS) tagging task using the Penn Treebank Wall Street Journal data. Additional details regarding these datasets are provided in Table 3.1. Note that tasks (2), (3), (4) are about language understanding. They use sparse features (the ratio a/A from the table is small). The sparsest data set is NER. Note that POS is considerably larger than other datasets. All experiments are performed with a regularization factor $\lambda = 1/n$. We used our own implementation[3] of SDCA coded in plain Python and Numpy (Walt et al., 2011). In most plots, we report the logarithm base 10 of the primal sub-optimality. We got the optimum by running L-BFGS for a large number of iterations.

3.6.2 Influence of the Line Search

We implemented the safe bounded Newton-Raphson method from Press et al. (1992, Section 9.4) on the derivative of the line search function. A natural question to ask is: how precise should the line search be? The stopping criterion for this algorithm is the size of the last step taken, so there is no proper precision parameter. We refer to this stopping criterion for the line search as the sub-precision of SDCA.

We discovered experimentally that the convergence of SDCA is mostly independent of the sub-precision. On all datasets, if we ask 0.01 sub-precision or less, SDCA converges with the same rate. An explanation is that the accuracy of the optimization arises from iterates $\boldsymbol{\alpha}$ and $\hat{\alpha}(\hat{w}(\boldsymbol{\alpha}))$ getting closer to each other in the simplex with each iteration.

Reaching 0.01 or 0.001 takes, on average, two iterations. Each iteration of

[3]The code to reproduce our experiments is available at `https://remilepriol.github.io/research/sdca4crf.html`.

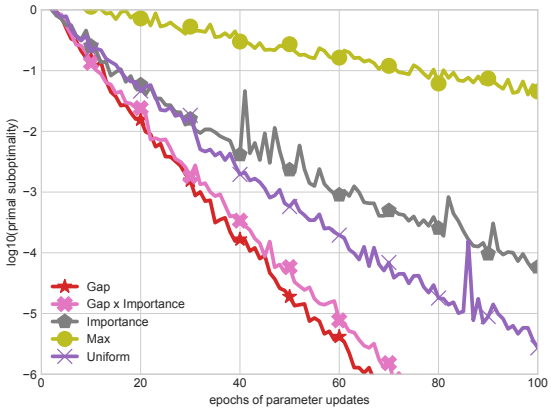


Figure 3.3 – Performance of competing sampling schemes on the OCR dataset with 80% of non-uniformity. Sampling proportionally to the gap gives the best performance.

Newton's method requires the computation of the first and second derivative of the line search objective (3.22). In the following, we report results with sub-precision 0.001 to be on the safe side. These two iterations were taking about 30% of the algorithms running time for each dataset.[4]

We also performed experiments with only one step of the Newton update. The convergence was not affected on OCR, CONLL, and POS, but convergence failed on NER (see Figure 3.9 of Appendix 3.E). This phenomenon could be related to sparsity.

3.6.3 Comparison of Sampling Schemes

We compare the performance of four sampling strategies with 20% of uniform sampling against the full Uniform approach, on the OCR dataset (see results in Figure 3.3):

- *Importance:* sample proportionally to the smoothness constants $L_i = \lambda + \frac{R_i}{n}$. We report how we evaluated the radii R_i in Appendix 3.C.
- *Gap:* sample proportionally to our current estimate of the duality gaps.[5]
- *Gap* $\times$ *importance:* sample proportionally to the product of the gap and smoothness constants.
- *Max:* sample deterministically the variable with the largest recorded gap (Dünner et al., 2017).

[4] We also tried initializing the line search with 0.5 or with the previous step size. There was no significant difference.

[5] For the gap approaches, we initialize the gap estimates with large values (100) so as to perform a pass over the whole dataset before starting to sample proportionally to the stale estimates.

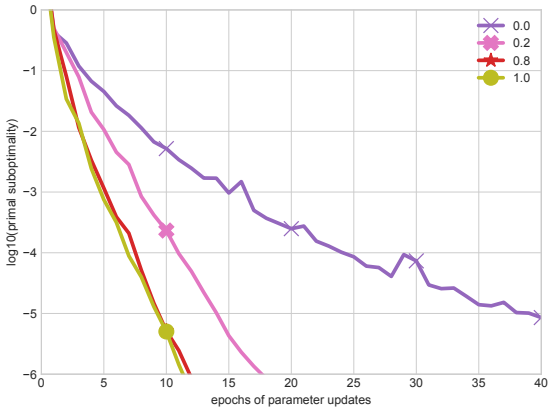
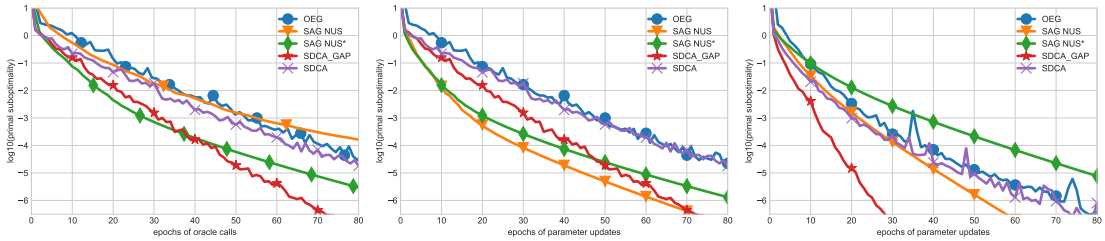


Figure 3.4 – SDCA with Gap sampling applied on NER with various fractions of non-uniform sampling, as indicated by the number in the legend. Increasing the fraction only improves the performance, up to a certain point.

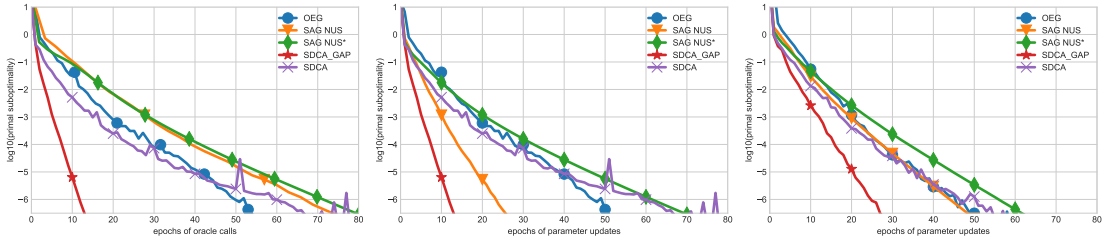
As discussed in Section 3.5.3, Max sampling is not robust enough to the staleness of the gap estimates and fails to converge here. We also observe that Importance performs worse than Uniform, and that Gap × Importance performs worse than Gap. This indicates that the smoothness upper bounds we estimated are not informative of the difficulty of optimizing a point for SDCA. Overall, Gap sampling gives the best performance, and this is what we use in the following experiments.

The ratio of uniform sampling is here to mitigate the fact that we sample proportionally to stale gaps. This is the strategy adopted by SAG-NUS (Schmidt et al., 2015) which samples uniformly half of the time. Another strategy used by Osokin et al. (2016) is to update all the duality gaps at once every ten epochs or so. Our experiments indicate that these strategies are not needed for SDCA-GAP. Increasing the ratio of non-uniformity up to 1 only improves the performance on all datasets, though after 0.8, the improvements are marginal, as illustrated by Figure 3.4 for the NER dataset.

In fact, the estimate of the total gap maintained by SDCA is somewhat accurate, as illustrated for different datasets in Figure 3.10 of Appendix 3.E. Empirically, it always remains within a factor 2 of the true duality gap. This accuracy is good news because one can use this estimate of the duality gap as a stopping criterion for the whole algorithm. Once it reaches a certain precision threshold, one just has to perform one last batch update to check the real value. This is similar in spirit to SAG, which uses the norm of its estimate of the true gradient as a stopping criterion. Both are duality gaps estimators (see Equation (3.9)).



(a) OCR (Oracle Calls) **(b)** OCR **(c)** CONLL



(d) NER (Oracle calls) **(e)** NER **(f)** POS

Figure 3.5 – Primal sub-optimality as a function of the number of oracle calls (left) or parameters updates (center and right). SDCA refers to uniform sampling. SDCA-GAP refers to sampling Gap sampling 80% of the time. SAG-NUS performs a line search at every iteration. SAG-NUS* implements a line-search skipping strategy. It appears worse than SAG-NUS when we look at the number of updates, which hides the cost of the line search.

3.6.4 Comparison against SAG and OEG

We downloaded the code for OEG and SAG-NUS as implemented by Schmidt et al. (2015) from the SAG4CRF project page.[6] We used our own implementation of SDCA with a line search sub-precision of 0.001. We provide the comparison in Figure 3.5 according to two different measures of complexity which are implementation independent.

Oracle calls. Schmidt et al. (2015) compared the algorithms on the basis of the number of oracle calls. We report these on OCR and NER in Figures 3.5a and 3.5d. Results on the other datasets are in Figure 3.7 in Appendix 3.E. This metric was suitable for the methods they compared. Both OEG and SAG-NUS use a line search where they call an oracle on each step. SDCA does not need the oracle to perform its line search. However, the oracle is a message passing on a junction tree. It has a cost proportional to the size of the marginals. Each iteration of the line search requires computing the entropy of these marginals, or their derivatives. These costs are roughly the same. Comparing the number of oracle calls for each method is thus unfairly advantaging SDCA by hiding the cost of its line search. It becomes a

[6] https://www.cs.ubc.ca/~schmidtm/Software/SAG4CRF.html

relevant comparison when a marginalization oracle becomes much more expensive than approximating the entropy (see the discussion in Section 3.7). When this cost is hidden, SDCA-GAP is on par with SAG-NUS* on OCR, and it is much faster on the sparse datasets.

Parameter updates. To give a different perspective, we report the log of the sub-optimality against the number of parameter updates in Figures 3.5b, 3.5c, 3.5e and 3.5f. This removes the additional cost of the line search for all methods.[7]

We observe that uniform SDCA and OEG need roughly the same number of parameters update on all four datasets. When we add the adaptive gap sampling, SDCA outperforms OEG by a margin. On OCR, SDCA and SDCA-GAP do not perform as well as SAG-NUS. On the three other datasets, SDCA-GAP needs fewer iterations. In fact, the more sparse the dataset, the fewer iterations are needed.

This is likely explained by SDCA's ability to almost perfectly optimize each block separately due to its line search method. More specifically, as the datasets become sparser, the prediction between data points becomes less and less correlated (i.e., the label distribution for two points that share no attributes will not influence each other directly through their primal weights). In settings where no points share any attributes (completely sparse), all methods optimize each point independently. SDCA may perform very well thanks to its precise line search.

In terms of test error, SDCA is on par with SAG, and a bit better than OEG. All methods reach maximum accuracy after a few epochs. We report the evolution of the test error in Figure 3.8 of Appendix 3.E.

Comparing the number of parameters updates also has a disadvantage. It penalizes methods with line search skipping strategies likes OEG and SAG. The running time is highly implementation dependent and providing a fair comparison is non-trivial. We focused on implementation-independent comparisons. SCDA, SAG, and OEG have many common operations: the oracle, the computation of the scores, and the primal direction. The fact that the line search took only 30% of SDCA's runtime indicates that the conclusion drawn from the number of updates may hold for other metrics.

3.7 Discussion

In this work, we investigated using SDCA for training CRFs for the first time. The observed empirical convergence per parameter update was similar for standard SDCA and OEG. However, SDCA can be enhanced with an adaptive sampling scheme, consistently accelerating its convergence and also yielding faster convergence than SAG with non-uniform sampling on datasets with sparse features. It would

[7] This is a penalty for SAG-NUS*, which enforces a line-search skipping strategy.

be natural to also implement a gap sampling scheme for OEG, though several quantities needed for the computation are not readily available in standard OEG and would yield higher overhead in actual implementation. We leave finding a more efficient implementation of a gap sampling scheme for OEG as an interesting research direction.

A key feature of SDCA is to only require one marginalization oracle per line search. This could become advantageous over SAG or OEG when the marginalization oracle becomes much more expensive than evaluating the entropy function from the marginals. Examples for this scenario include: when a parallel implementation is used for the entropy computation; or when the marginalization oracle uses an iterative approximate inference algorithm such as *tree reweighted belief propagation* whereas an approximation of the entropy is direct from the marginals (Krishnan et al., 2015). Investigating these scenarios with full timing comparison (which is implementation dependent) is a further interesting direction of future work.

We also note that acceleration schemes have been proposed for both SAG and SDCA (Lin et al., 2015; Shalev-Shwartz and Zhang, 2016), though they have not been tested yet for training CRFs.

3.A Implementation

We discuss some practical aspects of SDCA: initialization, memory requirement and how to do the line search.

3.A.1 Initialization

As discussed in Schmidt et al. (2015), the initialization of dual methods for CRFs can influence significantly their performance. We describe here a motivation for a suggested good initialization for $\boldsymbol{\alpha}$. Suppose that we put all the mass for α_i on the ground truth label y_i, i.e. $\alpha_i = \delta_{y_i}$ where δ_y is the Kronecker delta function on y – this represents the "empirical distribution" on one example. Let $\boldsymbol{\delta}$ be the concatenation $(\delta_{y_i})_{i=1}^n$. Similarly, let $\boldsymbol{u}$ be the concatenation of the uniform distribution on the labels for each training example. We have the following chain of relationships:

$$\begin{array}{ccccccc} \boldsymbol{\delta} & \xrightarrow{\hat{w}} & \mathbf{0} & \xrightarrow{\hat{\alpha}} & \boldsymbol{u} & \xrightarrow{\hat{w}} & \dots \\ \mathcal{D}(\boldsymbol{\delta}) = 0 & & \text{small } \mathcal{P}(\mathbf{0}) & & \mathcal{D}(\boldsymbol{u}) & & \end{array}$$

What is important here is that $\mathcal{P}(\mathbf{0})$ is small. If each node can take up to K values, and there are n sequences for a total of N nodes, $\mathcal{P}(\mathbf{0}) = \frac{N}{n}\log(K)$. On all our datasets this is below 100. This means that using $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\delta}$ gives an initial duality gap equal to $\mathcal{P}(\mathbf{0}) \lesssim 10^2$. In contrast, using $\boldsymbol{\alpha}^{(0)} = \boldsymbol{u}$ as used in the original OEG code[8] consistently gave extremely large $\hat{w}(\boldsymbol{u})$ resulting in a large negative dual score and large primal score, and raising numerical stability issues. Primal methods usually initialize their weights to zero. The dual counter part is the empirical distribution because it yields the same primal vector and score. For these reasons, we ideally would like to use $\boldsymbol{\delta}$ as the initialization.

There is catch though. On the borders of the simplex, the entropy has infinite gradient and curvature. This is a bad behavior if we wish to use this information for the line search. A natural strategy to mitigate this effect is to take a (small ϵ) convex combination with the uniform:

$$\boldsymbol{\alpha}^{(0)} := \varepsilon\boldsymbol{u} + (1-\varepsilon)\boldsymbol{\delta}\,. \tag{3.28}$$

This is what we use in our experiments. Graphically, the initial point will be on a segment between a corner of the simplex and the center. This is the same initialization that Schmidt et al. (2015, App. D of the Sup. Mat.) discovered empirically. It was also used implicitly by Collins et al. (2008) when they took the regularization path approach by starting the method with a very large regularization parameter λ.

[8] `egstra-0.2` available online at `http://groups.csail.mit.edu/nlp/egstra/`. This is also the initialization used in the main text of Schmidt et al. (2015).

3.A.2 Memory Requirement

Variance reduced methods use memory (except SVRG) to control the variance of the update. This memory cost can be quite large as it grows linearly with the size of the dataset. Schmidt et al. (2015) suggested a smart way to reduce this memory cost for SAG : for a sequence with hand crafted features, one stores only the unary marginals and the binary features. There is no such trick for dual methods, and both OEG and SDCA have to store the full marginals. It turns out that if each node can take K values, we have to allocate about K times more memory than for SAG. This can become a problem: for our larger dataset, part of speech tagging on Penn-Tree Bank Wall-Street Journal, we needed about 15GiB of RAM.

3.A.3 Line Search

The line search is an important part of the algorithm. Each evaluation of the line search function or its derivatives is quite expensive. We need to aggregate values from the whole marginal which has a size $\sum_c |\mathcal{Y}_c|$ (though this can be done in parallel). As a comparison, running the sum-product algorithm over the junction tree has a cost $2\sum_c |\mathcal{Y}_c|$ (though this is a sequential algorithm). There are other overhead in the algorithm such as computing the scores $\boldsymbol{w}^T F_c(x, y_c)$ or estimating the primal direction $A_i\delta_i$, so this is not totally critical.

Yet we wish to reduce the number of function evaluation. A good way to do so is to use the Newton-Raphson algorithm. But this uses the first and second derivatives of the line search objective, and the entropy has infinite slope and curvature on the borders of the simplex. To avoid numerical instability issues, we have to use and store the logarithm of the marginals (as was done for OEG (Collins et al., 2008)). We report an empirical study of the line search performance in section 3.6.2.

3.B Description of the Feature Map F

The feature map has the same structure on all the data sets (cf Figure 3.6). We first draw the distinction between unary features (in red) and binary features (in yellow). The features can be written as the sum of the unary and binary features:

$$F(x, y) = \sum_{t=1}^{T} F_t(x_t, y_t) + \sum_{t=1}^{T-1} F_{t,t+1}(y_t, y_{t+1}).$$

Unary Features depend only on the label of one node y_t and the corresponding data point x_t: $F_t(x_t, y_t)$. Binary features depend only on the labels of two neighboring nodes : $F_{t,t+1}(y_t, y_{t+1})$. It is a design choice not to directly model the relationship between two neighboring data points, e.g. $F(x_t, x_{t+1}, y_t, y_{t+1})$. In practice the

binary features simply count the number of transitions between y_t and y_{t+1}, hence the yellow square.

For unary features, it is a bit more complex. For each data sequence x, we extract an embedding for each position t, $\varphi(x,t)$. For OCR, it is simply the 128 pixels image itself $\varphi(x,t) = x_t$. For the language tasks, it is a count of the appearance of certain attributes, e.g, what is the word x_t, what are the words at position $t-1$, $t+1$, and so on. A complete list of the attributes is available at `http://www.chokkan.org/software/crfsuite/tutorial.html`. For each word (=node), between 13 and 20 features are extracted depending on the dataset. In total the number of different attributes extracted ranges from $73,000$ to $300,000$, hence the sparsity of the features. We denote A the number of attributes, or alternatively the size of the embedding. For each node with point x_t and label y_t, $F_t(x_t, y_t)$ puts the embedding $\varphi(x,t)$ in the column indexed by y_t of the red emission matrix. In this same column, we add some bias. The bias part has 3 dimensions. The first component counts the appearance of the label. The second component counts the appearance of the label in first position of a sequence, $(t=0)$. The last component counts the number of appearance in the last position of a sequence.

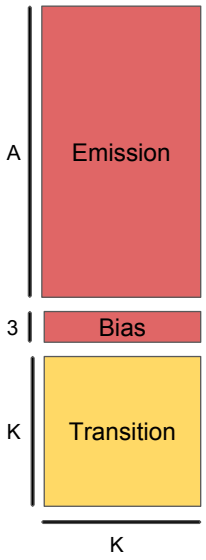


Figure 3.6 – Sketch of the feature map. K is the number of different labels for one node. A is the number of attributes.

3.C How to Compute the Radius of the Features

We drop the i index for now. We look at the pair (x,y). We want to evaluate an upper bound on:

$$R = \|A\|_{1\to 2}^2 = \max_{y\in\mathcal{Y}} \|\psi(y)\|_2^2 = \max_{\tilde{y}\in\mathcal{Y}} \|F(x,y) - F(x,\tilde{y})\|_2^2. \tag{3.29}$$

We are using the special nature of the features to estimate this radius. Remark that in the standard feature maps that we used (Appendix 3.B), there is one column per label. If the label y_t is assigned to the node t, then all the features extracted from that node are inserted in the column associated to y_t.

How to build a $\tilde{y}$ maximizing the distance between features? First we build the ground truth features : $F(x,y)$. Then we look at the labels included in the sequence y_t. In each data set, the K labels never appear together in one sequence.

We find a label z that does not appear in the original sequence. Then a sequence $\tilde{y}$ maximizing the objective (3.29) is the sequence composed only with that label z.

Why? There are two reasons. First, $F(x, y) \geq 0 \, \forall (x, y)$ thus we want $F(x, y)$ and $F(x, \tilde{y})$ to have disjoint supports such that the radius can be written as:

$$R = \|F(x, y)\|^2 + \|F(x, \tilde{y})\|^2. \tag{3.30}$$

Second, we want to maximize $\|F(x, \tilde{y})\|^2$. We need to put all the weights on few coordinates, instead of dispersing it. This is because we look at the ℓ^2 norm. For the ℓ^1 norm there would be no difference. By repeating only one label, we effectively concentrate all the weights in one column.

Following the steps described above, we can evaluate the radii for the whole data set.

3.D A Convergence Rate on the Duality Gap

It turns out that any algorithm with an upper bound on the primal or the dual sub-optimality for problems (3.2) and (3.3), can get an upper bound on the duality gap for the cost of a constant. To transpose a result of the *primal* sub-optimality to the duality gap, one can go by the norm of the gradient using the smoothness of $\mathcal{P}$, that we denote L:

$$\mathcal{P}(\boldsymbol{w}) - \mathcal{P}(\boldsymbol{w}^*) \geq \frac{1}{2L} \|\nabla \mathcal{P}(\boldsymbol{w})\|^2 \overset{(3.10)}{=} \frac{\lambda}{L} g(\boldsymbol{w}, \hat{\alpha}(\boldsymbol{w})). \tag{3.31}$$

The first inequality above is a standard one from convex analysis for convex functions with Lipschitz-continuous gradients (see e.g. (Nesterov, 2004b, eq. (2.1.6))). Whatever bound we get on the primal sub-optimality, we can translate it to the duality gap by losing a constant $L/\lambda \geq \kappa$, where κ is the condition number.

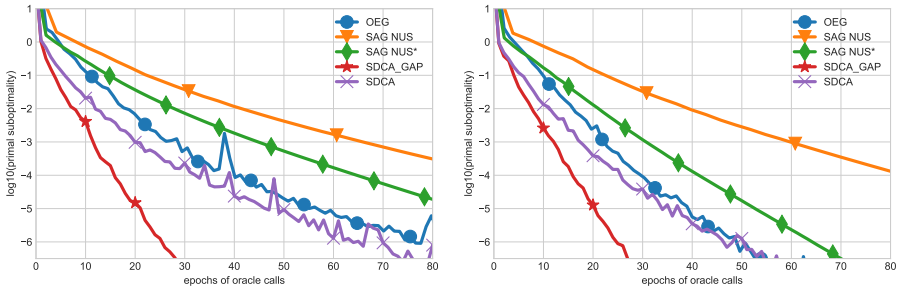
To transpose a result from the *dual* sub-optimality to the duality gap, one can use the uniform ascent lemma, Equation (3.74) from Appendix 3.F.4:

$$\mathcal{D}(\boldsymbol{\alpha}^*) - \mathcal{D}(\boldsymbol{\alpha}) \geq \mathbb{E}[\mathcal{D}(\boldsymbol{\alpha}^+)] - \mathcal{D}(\boldsymbol{\alpha}) \geq \frac{s}{n} g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) \tag{3.32}$$

where the expectation is taken over the stochasticity of the update. Let us look at this new constant. We know that $1/s = 1 + \frac{R}{n\lambda\mu}$. We can relate it to the smoothness $L \approx \lambda + \frac{R}{\mu}$. This time we lose a factor $n/s \approx n + \frac{L}{\lambda} \geq n + \kappa$. For a well-conditioned problem ($n \gg \kappa$) this is much larger than the constant we lose from the primal to the gap.

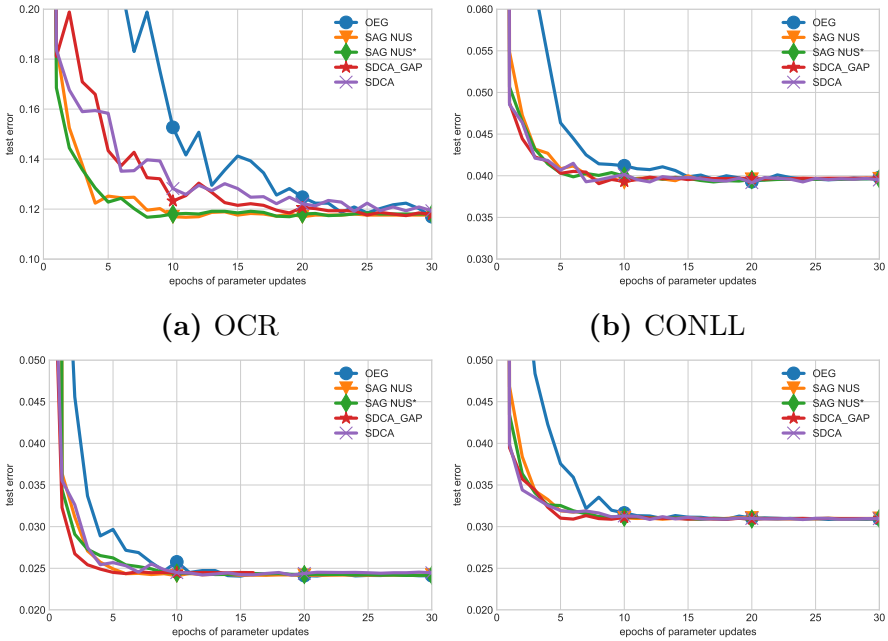
3.E Additional Comparison Plots

We provide additional figures on the primal sub-optimality as a function of oracle calls (Figure 3.7), the test error as a function of epochs (Figure 3.8), the impact of reducing the precision of the Newton line-search (Figure 3.9) and the ratio between the estimate of the duality gap and the ground truth (Figure 3.10).



(a) CONLL **(b)** POS

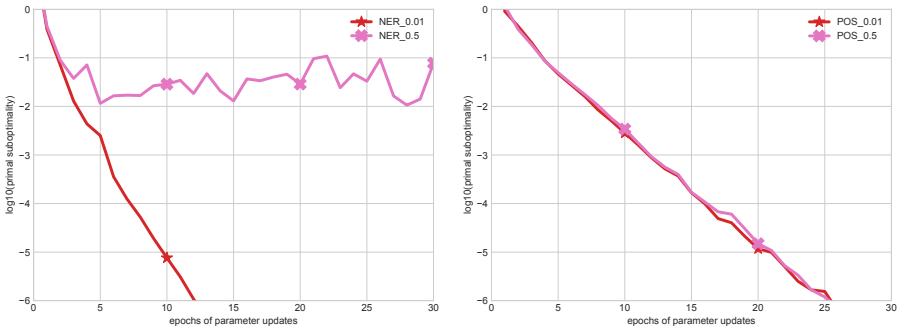
Figure 3.7 – Primal sub-optimality as a function of the number of oracle calls. SDCA-GAP performs much better than the competing methods for this metric partly because its line search does not require oracle calls.



(a) OCR **(b)** CONLL

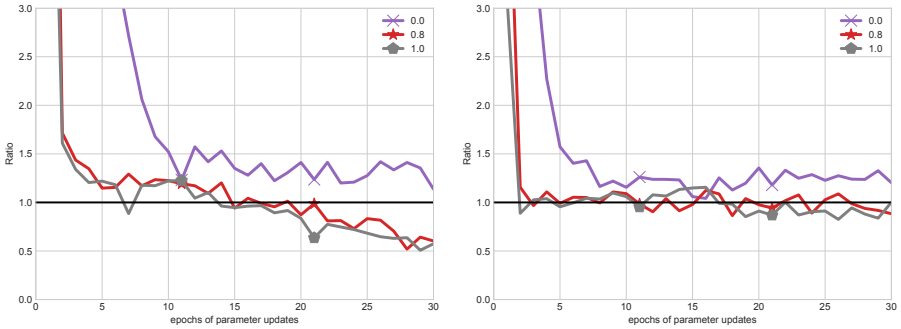
(c) NER **(d)** POS

Figure 3.8 – Test error against number of epochs. Every methods reach the same test error. SDCA and SAG have the same convergence speed.



(a) NER **(b)** POS

Figure 3.9 – Performance of SDCA on NER and POS with a Newton line-search. The number after the name of the dataset indicates the sub-precision we asked. A sub-precision of 0.5 effectively means that Newton stops after 1 step. While there is no difference between the curves for POS, 1 step of Newton update fails to converge on NER.



(a) CONLL **(b)** OCR

Figure 3.10 – The ratio between the estimate of the duality gap and the ground truth as a function of the proportion of non uniform sampling. The gap sampling tends to underestimate this value, whereas the uniform sampling tends to over-estimate it.

3.F A Technical Report on Non-uniform Sampling for Stochastic Dual Coordinate Ascent

In this section, we review the proofs of convergence of SDCA and its variants with importance and residual sampling. Then we derive bounds on the convergence rate of two new sampling scheme for SDCA. The first scheme samples proportionally to the duality gaps of each individual variable. The second scheme is similar to the

first one, but it corrects the duality gaps with the Lipschitz constant of the primal problem.

3.F.1 Setting

We derive these bounds in a more general setting than the logistic regression, and we have to introduce some new notation.

Let $\boldsymbol{w}$ denote the weights vector parameter, and A_i the i-th features matrix. Let ϕ be the primal loss function. We suppose it is convex and $1/\mu$-smooth with respect to $\|.\|_P$ (dual norm $\|.\|_D$). The regularizer r is supposed 1-strongly convex with respect to $\|.\|_{P'}$ (dual norm $\|.\|_{D'}$). Because ϕ and r^* are smooth, they are also differentiable. Note that every starred variable represent its dual conjugate.

The empirical loss minimization problem is:

$$(P) \quad \min_{\boldsymbol{w}\in\mathbb{R}^d} \lambda r(\boldsymbol{w}) + \frac{1}{n}\sum_{i=1}^{n} \phi_i(-A_i^T\boldsymbol{w}). \tag{3.33}$$

Its Fenchel dual problem is:

$$(D) \quad \max_{\boldsymbol{\alpha}|\forall i,\alpha_i\in \mathrm{Dom}\,\phi^*} -\lambda r^*(\hat{v}(\boldsymbol{\alpha})) - \frac{1}{n}\sum_{i=1}^{n} \phi_i^*(\alpha_i), \tag{3.34}$$

with

$$\hat{v}(\boldsymbol{\alpha}) := \frac{1}{\lambda n}\sum_i A_i\alpha_i \quad \text{and} \quad \hat{w}(\boldsymbol{\alpha}) \in \nabla r^*(\hat{v}(\boldsymbol{\alpha})). \tag{3.35}$$

We also note:

$$\forall i, \beta_i = \hat{\alpha}_i(\boldsymbol{w}) \in \nabla\phi_i(-A_i^T\boldsymbol{w}). \tag{3.36}$$

Minimization of the empirical risk can often be interpreted as going around the diagram below.

$$\begin{array}{ccc} \boldsymbol{w} & \longrightarrow & \nabla\phi_i(-A_i^T\boldsymbol{w}) \\ \uparrow & & \downarrow \\ \nabla r^*\left(\frac{1}{\lambda n}A\boldsymbol{\alpha}\right) & \longleftarrow & \boldsymbol{\alpha} \end{array}$$

We define the squared radius of the features for a given sample i as the operator norm of the matrix A_i:

$$R_i := \|A_i\|^2_{D\to D'}. \tag{3.37}$$

We also define the maximum squared radius as $R = \max_i R_i$ and the mean radius $\bar{R} = \frac{1}{n}\sum_i R_i$.

Log-likelihood special case. The loss $\phi(z) = \log(\sum_y \exp(z_y))$ is 1-smooth with respect to the max-norm. Its convex conjugate is the negative entropy $\phi^*(\alpha) = -H(\alpha) = \sum_y \log(\alpha_y)\alpha_y$ which is in turn 1-strongly convex with respect to the ℓ_1-norm, and whose domain is the simplex. We use the ℓ_2 regularization whose dual function is itself. We thus have $R_i = \|A_i\|_{1\to 2}^2 = \max_y \|\psi_i(y)\|_2^2$. We also have a special expression for the primal to dual function $\beta_i = p(.|x_i; \boldsymbol{w}) \propto \exp(-\boldsymbol{w}^T \psi_i(.))$. The dual variable is obtained as the conditional probability of the primal model. Conversely, the primal weights are obtained as the expectation of the features $\psi_i(y)$, which are the columns of A_i.

3.F.2 Duality Gaps

We derive an interesting form on the duality gaps that support a new sampling strategy. This is not needed to understand the convergence rates of SDCA and its variants, and the reader may skip this section.

The duality gap is:

$$g(\boldsymbol{w}, \boldsymbol{\alpha}) = P(\boldsymbol{w}) - D(\boldsymbol{\alpha}) = \lambda \left(r(\boldsymbol{w}) + r^*(\frac{A\boldsymbol{\alpha}}{\lambda n}) \right) + \frac{1}{n} \sum_{i=1}^n \phi(-A_i^T \boldsymbol{w}) + \phi^*(\alpha_i). \quad (3.38)$$

Because of the two conjugate pairs (r, r^*) and (ϕ, ϕ^*) there are two apparent ways to simplify it. One is to take the conjugate primal variable $\boldsymbol{w} := \hat{w}(\boldsymbol{\alpha})$, another is to take the conjugate dual variable $\boldsymbol{\alpha} := \hat{\alpha}(\boldsymbol{w})$.

Conjugate primal variable. Under the hypothesis $\boldsymbol{w} = \hat{w}(\boldsymbol{\alpha})$, we obtain:

$$r(\boldsymbol{w}) + r^*(\frac{A\boldsymbol{\alpha}}{\lambda n}) = \boldsymbol{w}^T \frac{A\boldsymbol{\alpha}}{\lambda n}. \quad (3.39)$$

The duality gap simplifies:

$$g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \frac{1}{n} \sum_{i=1}^n \phi(-A_i^T \hat{w}(\boldsymbol{\alpha})) + \phi^*(\alpha_i) - \alpha_i^T(-A_i^T \hat{w}(\boldsymbol{\alpha})) = \frac{1}{n} \sum_{i=1}^n F_\phi(-A_i^T \hat{w}(\boldsymbol{\alpha}), \alpha_i), \quad (3.40)$$

where $F_\phi(s, \alpha)$ is the Fenchel duality gap (3.11) between vectors s and α. When ϕ is the log-sum-exp, these vectors are the score (or logit) s and the probability α. We want to simplify this further to directly relate $\boldsymbol{\alpha}$ and its next iterate $\hat{\alpha}_i \circ \hat{w}(\boldsymbol{\alpha})$. To do so we need another condition:

$$\langle \nabla\phi^* \circ \nabla\phi(s) - s, \beta - \alpha \rangle = 0, \quad (3.41)$$

for all $s \in \operatorname{Dom} \phi$ and $\alpha, \beta \in \operatorname{Dom} \phi^*$. Geometrically, the pairs $(s, \nabla\phi^* \circ \nabla\phi(s))$ should always be aligned orthogonally to $\operatorname{Dom} \phi^*$. This condition (3.41) is true

whenever $\nabla\phi^* \circ \nabla\phi = \mathrm{Id}$ the identity function. It is also true when ϕ is the log-sum-exp although $\nabla\phi^* \circ \nabla\phi$ is not the identity. Then the Fenchel duality gap is equal to the Bregman divergence generated by ϕ^*:

$$F_\phi(s, \alpha) = D_{\phi^*}(\alpha || \nabla\phi(s)). \tag{3.42}$$

Then the duality gap can be written as the average over data points of the ϕ^*-Bregman divergence between α_i and its next fixed point iterate: $\hat{\alpha}_i \circ \hat{w}(\boldsymbol{\alpha})$:

$$g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} D_{\phi^*}(\alpha_i || \hat{\alpha}_i \circ \hat{w}(\boldsymbol{\alpha})). \tag{3.43}$$

Conjugate dual variable. The situation is quite symmetric. Under the assumption that $\boldsymbol{\alpha} := \hat{\alpha}(\boldsymbol{w})$, one gets:

$$g(\boldsymbol{w}, \hat{\alpha}(\boldsymbol{w})) = \lambda \left(r(\boldsymbol{w}) + r^*(\frac{A\hat{\alpha}(\boldsymbol{w})}{\lambda n})) - \boldsymbol{w}^T \frac{A\hat{\alpha}(\boldsymbol{w})}{\lambda n} \right) = \lambda F_r(\boldsymbol{w}, \frac{A\hat{\alpha}(\boldsymbol{w})}{\lambda n}), \tag{3.44}$$

where F_r is the fenchel duality gap of the regularizer. We can transform it into the Bregman divergence between $\boldsymbol{w}$ and its next iterate $\boldsymbol{w}' := \nabla r^*(\frac{A\hat{\alpha}(\boldsymbol{w})}{\lambda n})) = \hat{w} \circ \hat{\alpha}(\boldsymbol{w})$ at the condition that:

$$\langle \nabla r \circ \nabla r^*(\boldsymbol{v}) - \boldsymbol{v}, \boldsymbol{w}' - \boldsymbol{w} \rangle = 0, \tag{3.45}$$

for all vectors $\boldsymbol{v}$ in the domain of r^* and all vectors $\boldsymbol{w}, \boldsymbol{w}'$ in the domain of r. Then the duality gap is:

$$g(\boldsymbol{w}, \hat{\alpha}(\boldsymbol{w})) = \lambda D_r(\boldsymbol{w} || \hat{w} \circ \hat{\alpha}(\boldsymbol{w})). \tag{3.46}$$

Equations (3.43) and (3.46) show that the objective (3.33) is also a fixed point problem for the conjugation operations. The suboptimality can be easily measured as the divergence between a point, either primal or dual and its next iterate. The divergence is given by the regularizer of the primal problem r or the dual problem ϕ^*.

3.F.3 Theorems

We state the convergence rates for some variants of SDCA using non-uniform sampling. The proofs follow in the next section.

Denote $h_t := D(\boldsymbol{\alpha}^*) - \mathbb{E}[D(\boldsymbol{\alpha}^{(t)}]$ the expectation of the dual sub-optimality at step t. The expectation is over all the possible samplings (the stochastic part of SDCA). We will bound this value. One can bound the duality gap $g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) := P(\hat{w}(\boldsymbol{\alpha})) - D(\boldsymbol{\alpha})$ at the cost of another constant outside of the exponential (Appendix 3.D).

Theorem 3.F.1 (Uniform sampling (Shalev-Shwartz and Zhang, 2013a)). *At each step, sample i with uniform probability in $[1, n]$. After t iterations, the dual sub-optimality is bounded by:*

$$h_t \leq (1 - \frac{s}{n})^t h_0, \tag{3.47}$$

where $s = (1 + \frac{R}{n\lambda\mu})^{-1}$ is the fixed step size used in the proof.

This theorem holds for SDCA with line search as well, since the line search can only be faster than the fixed step size. None of the following algorithm take the line search into account. The relative values of the bounds appearing in each theorems may not always reflect the relative performance of each algorithms.

Intuitively, we want the linear coefficient, here $\frac{s}{n}$, to be as large as possible. Here R/μ is the max of the smoothness of the individual losses ϕ_i. If the regularizer is smooth enough, then the linear coefficient is related to the condition number κ by:

$$\frac{n}{s} = n + R/(\lambda\mu) \approx n + \kappa. \tag{3.48}$$

The following theorem goes from the maximum radius R to the mean radius $\bar{R}$.

Theorem 3.F.2 (Importance Sampling (Zhao and Zhang, 2015)). *At each step, sample i with probability p_i proportional to the individual "condition number":*

$$p_i \propto 1 + R_i/(n\lambda\mu). \tag{3.49}$$

After t iterations, the dual sub-optimality is bounded by:

$$h_t \leq (1 - \frac{\bar{s}}{n})^t h_0, \tag{3.50}$$

where $\bar{s} := (1 + \frac{\bar{R}}{n\lambda\mu})^{-1}$ is the harmonic mean of the step sizes used in the proof.

The harmonic mean is always larger than the minimum step size, so the importance sampling will converge faster than the uniform sampling *at the condition* that we have an accurate estimate of the operator norms R_i. Indeed, if we get the operator norms wrong, then we will sample more often points that are actually easier to classify. Even if we estimate them right, empirical convergence may be slower with this scheme because of the line search. This is what happened during the experiments that we ran on CRFs.

Note the similarity with non-uniform sampling in primal methods. The convergence is improved thanks to larger step sizes, that are proportional to the inverse of some kind of Lipschitz constants. The convergence rate depends on the arithmetic mean of these Lipschitz constants instead of the max.

We now introduce an adaptive scheme. We reformulate the theorem to make it more compact and comparable with our theorems.

Theorem 3.F.3 (AdaSDCA (Csiba et al., 2015))**.** *Suppose that the loss functions are* ***quadratic*** $\phi(z) := \|z\|_2^2$. *Denote* $d_i^t = \|\beta_i^t - \boldsymbol{\alpha}_i^t\|_{D'}$ *At each step* t*, sample* i *with probability* p_i^t *defined by:*

$$p_i^t \propto d_i^t \sqrt{1 + R_i/(n\lambda\mu)}, \tag{3.51}$$

$$\theta(\boldsymbol{d}, \boldsymbol{p}) = \frac{\sum_i d_i^2}{\sum_{i|p_i>0} \frac{d_i^2}{p_i}(1 + \frac{R_i}{n\lambda\mu})}, \tag{3.52}$$

and

$$\tilde{\theta}_t = \frac{\mathbb{E}[\theta(\boldsymbol{d}^t, \boldsymbol{p}^t)(P(\boldsymbol{w}^t) - D(\boldsymbol{\alpha}^t))]}{\mathbb{E}[P(\boldsymbol{w}^t) - D(\boldsymbol{\alpha}^t)]} \tag{3.53}$$

where the expectation is taken over all the possible trajectories of the algorithm, e.g the sampling of the points. Finally define $\tilde{\theta} = \min_t \tilde{\theta}_t$*. After* t *iterations, the dual sub-optimality is bounded by:*

$$h_t \leq (1 - \tilde{\theta})^t h_0. \tag{3.54}$$

In the theorem above, we have to take the expectation of some variable over all the trajectories of the algorithm. This is not very clean, but this is unavoidable to get a general convergence result with an adaptive scheme. Alternatively, one could simply compare the improvement given by one step for each algorithm.

A major limitation of the theorem above is that the loss has to be quadratic. This theoretical limitation is not a big problem empirically. It results from a symbolic trick used in the proof : setting the step size to be proportional to the inverse of the probability. This is reasonable for importance sampling, because the probability is proportional to the smoothness constant. Setting the step size to the inverse of the smoothness is optimal for gradient descent. This may be less reasonable for other sampling schemes.

Another limitation is that we have to estimate the n distances d_i^t at each step. In practice we compute d_i^t only for the sampled i, and use the latest estimate $d_j^{t'}$ for all the other samples j. Our estimates will become stale as the algorithm unfolds, but there are heuristics to compensate for that phenomenon. One is to sample from a mixture between a uniform and an adaptive distribution. Another is to do a batch update of the d_i every once in a while. These heuristics are unavoidable for adaptive schemes, as we do not want the cost of every update to be $O(n)$. We do not know how to analyze the impact of these heuristics. Empirically, adaptive sampling with this heuristic still accelerates convergence.

Now we are going to introduce two new adaptive sampling scheme. Both of them rely on the structure of the duality gap:

$$g(\hat{w}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) := P(\hat{w}(\boldsymbol{\alpha})) - D(\boldsymbol{\alpha}) = \sum_i \phi(-A_i^T \hat{w}(\boldsymbol{\alpha})) + \phi^*(\alpha_i) + \langle \hat{w}(\boldsymbol{\alpha}), A_i \alpha_i \rangle. \tag{3.55}$$

Each term of the sum above is a Fenchel duality gap between the loss and its convex conjugate. They are all positive, and somehow represent the sub-optimality of the current model for every training sample. Intuitively, sampling the most sub-optimal point may yield the best improvement.

Theorem 3.F.4 (Gap sampling)**.** *At each step t, sample i with probability p_i^t proportional to the individual Fenchel duality gap:*

$$p_i^t \propto g_i^t := \phi(-A_i^T \boldsymbol{w}^t) + \phi^*(\alpha_i^t) + \langle \boldsymbol{w}^t, A_i \alpha_i^t \rangle. \tag{3.56}$$

Define the non-uniformity of the duality gaps as the ratio between their quadratic mean and their arithmetic mean:

$$\chi^2(\boldsymbol{g}) := \frac{\frac{1}{n}\sum_i g_i^2}{\left(\frac{1}{n}\sum_i g_i\right)^2} \in [1, n]. \tag{3.57}$$

Take χ a lower bound on these non-uniformity over all trajectories, for all time steps. After t iterations, the dual sub-optimality is bounded by:

$$h_t \leq (1 - s\frac{\chi^2}{n})^t h_0. \tag{3.58}$$

where $s = (1 + \frac{R}{n\lambda\mu})^{-1}$ is the fixed step size used in the proof.

This theorem has the same limitations relative to adaptive scheme that we mentioned for AdaSDCA.

This kind of sampling scheme was studied in the sublinear convergence regime by Osokin et al. (2016) (Franke-Wolfe) and Perekrestenko et al. (2017) (Coordinate Descent). They could not establish a domination of gap-sampling over uniform sampling. This is what we prove in the linear regime for SDCA since the non-uniformity χ belongs to $[1, \sqrt{n}]$.

The non-uniformity $\chi^2(\boldsymbol{g})$ (3.57) is worth 1 if the gaps are all the same, and $\sqrt{n}$ if only one gap is non-zero, hence the name. Gap-sampling will be n times faster than uniform sampling if only one sample i is suboptimal $g_i > 0$. This result is sensible since we will sample only one point, while the uniform algorithm may sample a large number first. Let us imagine another scenario where all points are already optimal except k of them which have the same gap value. Then the acceleration coefficient will be $\frac{n}{k}$, which can be a significant acceleration when k is much smaller than n. Finally, consider a scenario where the gaps are evenly distributed $\{a, 2a, ..., na\}$ for some value $a > 0$. Note that $\chi^2(\boldsymbol{g})$ is scale-invariant and does not depend on the specific value a. We can compute $\chi^2(\boldsymbol{g})$ explicitly here using Faulhaber's formula for the sum of powers of integers:

$$\chi^2(\boldsymbol{g}) = \frac{\frac{1}{n}\frac{n(n+1)(2n+1)}{6}}{\left(\frac{1}{n}\frac{n(n+1)}{2}\right)^2} = \frac{2}{3}\frac{2n+1}{n+1} \approx 4/3.$$

The acceleration coefficient here is approximately 4/3 compared to uniform sampling.

The duality gaps are often computable, even in the Conditional Random Fields context. On the other hand, we do not have direct access to the dual variable $\boldsymbol{\alpha}$ and we cannot compute the distance $d_i = \|\beta_i - \alpha_i\|_1$, as it is the ℓ^1 norm of a vector of exponential size.

Now we want to combine importance sampling with duality gap sampling. We would like to benefit both from the dependency on $\bar{R}$ and the acceleration by χ.

Theorem 3.F.5 (Lipschitz-gap sampling)**.** *At each step t, sample i with probability p_i^t defined by:*

$$p_i^t \propto g_i^t(1 + R_i/(n\lambda\mu)). \tag{3.59}$$

Define χ as in (3.57) from Theorem 3.F.4. Define $\tilde{s}$ as the quadratic harmonic mean of the step sizes $s_i := 1/(1 + R_i/(n\lambda\mu))$. After t iterations, the dual sub-optimality is bounded by:

$$h_t \leq (1 - \tilde{s}\frac{\chi}{n})^t h_0. \tag{3.60}$$

This theorem makes apparent a trade-off between the advantage gained with the smoothness, and the advantage gained with the individual gaps. We lose the square factor on the non-uniformity compared to Theorem 3.F.4. We go from the harmonic mean to the quadratic harmonic mean (generalized norm -2) of the step sizes, which is basically the same as going from the arithmetic mean of the smoothness to the quadratic mean of the smoothness. Recall that the quadratic mean always lies in between the arithmetic mean and the max.

Our results holds for any smooth loss function, contrary to AdaSDCA. Our two new strategies complement importance sampling as none of them dominates the other. Which one is the best depends on the context. *That is* at the condition that we have access to the R_i. Otherwise gap sampling remains available.

3.F.4 Proofs

Lemma 3.F.6 (General descent lemma)**.** *Apply the SDCA update on the dual variable $\boldsymbol{\alpha}$ to get the new point $\boldsymbol{\alpha}^+$. The block i is sampled with probability p_i and updated with a step size s_i. The expected dual improvement verifies the lower bound:*

$$n\mathbb{E}_{\boldsymbol{p}}[D(\boldsymbol{\alpha}^+)] - D(\boldsymbol{\alpha}) \geq \underbrace{\sum_i p_i s_i g_i}_{\text{not the duality gap}} + \frac{\mu}{2}\sum_i p_i s_i \Bigg(1 - s_i \underbrace{\left(1 + \frac{R_i}{\mu\lambda n}\right)}_{:=c_i}\Bigg) d_i^2 \tag{3.61}$$

where $\mathbb{E}_{\boldsymbol{p}}$ denotes the conditional expectation over the choice $i \sim \boldsymbol{p}$ of block to update, conditioned on the previous state $\boldsymbol{\alpha}$.

Proof of Lemma 3.F.6. This statement is similar to a weighted combination of Equation (25) from Shalev-Shwartz and Zhang (2013a). We provide here the

derivation to be self-contained. Suppose we sampled the point i and updated the block α_i with step size s_i:

$$\alpha_i^+ := \alpha_i + s_i\delta_i = (1-s_i)\alpha_i + s_i\beta_i\,. \tag{3.62}$$

The dual improvement is:

$$n(D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})) = \underbrace{\lambda n\left(r^*\left(\frac{A\boldsymbol{\alpha}}{\lambda n}\right) - r^*\left(\frac{A\boldsymbol{\alpha}^+}{\lambda n}\right)\right)}_{\text{data fidelity}} + \underbrace{\phi^*(\alpha_i) - \phi^*(\alpha_i^+)}_{\text{regularization}}\,. \tag{3.63}$$

We first bound the data fidelity term. We use the the fact that r^* is 1-smooth with respect to $\|.\|_{D'}$ to upper-bound its variation:

$$r^*\left(\frac{A\boldsymbol{\alpha}^+}{\lambda n}\right) = r^*\left(\frac{A\boldsymbol{\alpha}}{\lambda n} + s_i\frac{A_i\delta_i}{\lambda n}\right) \tag{3.64}$$

$$\leq r^*\left(\frac{A\boldsymbol{\alpha}}{\lambda n}\right) + s_i\left\langle \nabla r^*\left(\frac{A\boldsymbol{\alpha}}{\lambda n}\right), \frac{A_i\delta_i}{\lambda n}\right\rangle + \frac{s_i^2}{2}\left\|\frac{A_i\delta_i}{\lambda n}\right\|_{D'}^2 \tag{3.65}$$

The linear coeficient of this lower boudn is $\hat{w}(\boldsymbol{\alpha}) = \nabla r^*\left(\frac{A\boldsymbol{\alpha}}{\lambda n}\right)$. The quadratic term can be further upper-bounded:

$$\left\|\frac{A_i\delta_i}{\lambda n}\right\|_{D'}^2 \leq \frac{1}{(\lambda n)^2}\|A_i\|_{D\to D'}^2\|\delta_i\|_D^2 = \frac{R_i d_i^2}{(\lambda n)^2}\,, \tag{3.66}$$

by definition of the radius R_i and the residue $d_i := \|\beta_i - \alpha_i\|_D$. So the loss variation is lower bounded by:

$$\lambda n\left(r^*\left(\frac{A\boldsymbol{\alpha}}{\lambda n}\right) - r^*\left(\frac{A\boldsymbol{\alpha}^+}{\lambda n}\right)\right) \geq s_i\left\langle \hat{w}(\boldsymbol{\alpha}), A_i(\alpha_i - \beta_i)\right\rangle - \frac{s_i^2}{2}\frac{R_i d_i^2}{\lambda n}\,. \tag{3.67}$$

Now we bound the regularization term. Since ϕ^* is μ-strongly convex with respect to $\|.\|_D$,

$$\phi^*(\alpha_i^+) = \phi^*((1-s_i)\alpha_i + s_i\beta_i) \leq (1-s_i)\phi^*(\alpha_i) + s_i\phi^*(\beta_i) - s_i(1-s_i)\frac{\mu}{2}d_i^2\,. \tag{3.68}$$

The regularization variation can be lower bounded by:

$$\phi^*(\alpha_i) - \phi^*(\alpha_i^+) \geq s_i\left(\phi^*(\alpha_i) - \phi^*(\beta_i)\right) + s_i(1-s_i)\frac{\mu}{2}d_i^2\,. \tag{3.69}$$

Plugging the bounds (3.67) and (3.69) into Equation (3.63), we get:

$$n(D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha}))$$

$$\geq s_i\left(\phi^*(\alpha_i) + \left\langle \hat{w}(\boldsymbol{\alpha}), A_i(\alpha_i - \beta_i)\right\rangle - \phi^*(\beta_i)\right) + \frac{s_i}{2}\left((1-s_i)\mu - s_i\frac{R_i}{\lambda n}\right)d_i^2\,. \tag{3.70}$$

Recall that $\beta_i := \nabla\phi(-A_i^T\hat{w}(\boldsymbol{\alpha}))$. Thus,

$$\left\langle -A_i^T\hat{w}(\boldsymbol{\alpha}), \beta_i \right\rangle - \phi^*(\beta_i) = \phi(-A_i^T\hat{w}(\boldsymbol{\alpha})) \tag{3.71}$$

by definition of the convex conjugate ϕ^*. To sum up, at iteration t, if we sample the block i, and update it with step size s_i, we can lower bound the resulting dual improvement with:

$$\begin{aligned} &n(D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})) \\ &\geq s_i \underbrace{\left[\phi(-A_i^T\hat{w}(\boldsymbol{\alpha})) + \phi^*(\alpha_i) + \hat{w}(\boldsymbol{\alpha})^T A_i\alpha_i\right]}_{\text{Fenchel gap}=:g_i} + \frac{s_i\mu}{2}\left(1 - s_i\left(1 + \frac{R_i}{\mu\lambda n}\right)\right) d_i^2 \,. \end{aligned} \tag{3.72}$$

To conclude the proof, take a weighted average of the inequalities (3.72) with the weights p_i. □

In the following we note the duality gap:

$$\bar{g} := \frac{1}{n}\sum_i g_i = P(\hat{w}(\boldsymbol{\alpha})) - D(\boldsymbol{\alpha}) \,. \tag{3.73}$$

Proof of Theorem 3.F.1. In the original proof of Shalev-Shwartz and Zhang (2013a), we set $p_i = 1/n$ and $s_i = s = (1 + \frac{R}{n\lambda\mu})^{-1} \leq 1/c_i$. This step size guarantees that the right hand term is positive, leaving us with the inequality:

$$\mathbb{E}_{p^t}[D(\boldsymbol{\alpha}^{t+1}) - D(\boldsymbol{\alpha}^t)] \geq \frac{s}{n}\bar{g}^t. \tag{3.74}$$

Now observe that $\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] = -\mathbb{E}_p[h_{t+1}] + h_t$ and $\bar{g}^t = (P(\boldsymbol{w}^t) - D(\boldsymbol{\alpha}^t)) \geq h_t$. Moving the sub-optimality at time t on the right gives:

$$\mathbb{E}_p[h_{t+1}] \leq (1 - \frac{s}{n})h_t. \tag{3.75}$$

This inequality is conditional on all the random sampling until time t. Let us take the expectation of this inequality with respect to all this past randomness. We get a recursive upper bound on the expected dual sub-optimality:

$$\mathbb{E}[h_{t+1}] \leq (1 - \frac{s}{n})\mathbb{E}[h_t] \leq (1 - \frac{s}{n})^t h_0. \tag{3.76}$$

This is the final convergence result with the linear constant $s/n = (n + R/(\lambda\mu))^{-1}$. □

In the proof above, we lower bound the dual improvement by the duality gap, then we use this to get the linear convergence rate. All the proofs follow the same reasoning, and the last few steps are always the same so we will skip them.

Proof of Theorem 3.F.2. Inject $p_i = c_i / \sum_j c_j$ and $s_i = 1/c_i$. The right hand term is zero thanks to the step size, hence the lower bound:

$$\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{\bar{g}}{\sum_i c_i}. \tag{3.77}$$

We get the linear rate $\frac{1}{\sum_i c_i}$ which is also the harmonic mean of the step sizes divided by n. □

Sketch of Proof of Theorem 3.F.3. To make the duality gap appear in this formula for arbitrary probability p, Csiba et al. (2015) use $p_i s_i = \theta$ constant, whenever $g_i > 0$. If the individual duality gap is null $g_i = 0$, then they set $p_i = s_i = 0$.

$$\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] - \theta\bar{g} \geq \theta\frac{\mu}{2n}\sum_i d_i^2\left(1 - \frac{\theta}{p_i}\left(1 - \frac{R_i^2}{\mu\lambda n}\right)\right) \tag{3.78}$$

The negative consequence of that strategy is that they have to enforce $s_i \in [0, 1]$ by setting $\theta < \min_i p_i$ where the minimum is taken over the sub-optimal i's (i.e. $p_i > 0$). This a terrible constraint on the step size, as we cannot be too non-uniform without taking very small steps. It effectively reduces the linear convergence constant θ/n.

Finally, they want to maximize θ while keeping the right hand side positive. This is a hard problem on θ and p. When the loss is the quadratic loss, they can remove the condition that the step size should be smaller than 1. Then they solve the optimization problem to get the sampling scheme $p_i \propto d_i\sqrt{c_i}$. □

Proof of Theorem 3.F.4. We use the same step size as in the original proof:

$$s_i = s = \frac{n}{n + R/(\lambda\mu)}. \tag{3.79}$$

We have the guarantee that the right hand term is positive. The lemma simplifies to:

$$n\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{s}{n}\sum_i p_i g_i. \tag{3.80}$$

We inject $p_i = \frac{g_i}{n\bar{g}}$ into this lower bound:

$$\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{s}{n}\frac{\sum_i g_i^2}{\sum_j g_j} = \frac{s}{n}\chi^2(\boldsymbol{g})\bar{g}\,, \tag{3.81}$$

where we introduced the non-uniformity of the duality gaps vector defined in Equation (3.57). To get a simpler expression for a global convergence bound, let us define χ to be a lower bound on $\chi(\boldsymbol{g})$ over all the possible unfolding of SDCA and for every steps. Now we can write the descent lemma in the same form as in the original proof, but a with new constant:

$$\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{s}{n}\chi^2\bar{g}\,. \tag{3.82}$$

□

Proof of Theorem 3.F.5. We set $p_i \propto g_i c_i$ where $c_i = 1 + R/(n\lambda\mu)$.

$$n\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{\sum_i s_i g_i^2 c_i}{\sum_i g_i c_i} + \frac{\frac{\mu}{2}\sum_i s_i g_i c_i d_i^2 \left(1 - s_i c_i\right)}{\sum_i g_i c_i} \tag{3.83}$$

Similarly to the proof of importance sampling, we now set $s_i = 1/c_i \leq 1$ instead of $s_i = s = 1/\max_i c_i$. This nullifies the right hand term. We can take longer steps if the individual Lipschitz constants are high.

$$n\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{\sum_i g_i^2}{\sum_i g_i c_i} = \frac{\langle \boldsymbol{g}, \boldsymbol{g} \rangle}{\langle \boldsymbol{c}, \boldsymbol{g} \rangle} \tag{3.84}$$

We apply the Cauchy-Schwartz inequality : $\langle \boldsymbol{c}, \boldsymbol{g} \rangle \leq \|\boldsymbol{c}\|_2 \|\boldsymbol{g}\|_2$.

$$n\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{\|g\|_2}{\|c\|_2} = \frac{\chi(g)}{QM(c)}\bar{g}\,, \tag{3.85}$$

where QM denotes the quadratic mean. Finally we divide both sides by n to complete the proof:

$$\mathbb{E}_p[D(\boldsymbol{\alpha}^+) - D(\boldsymbol{\alpha})] \geq \frac{\chi(g)}{nQM(c)}\bar{g}\,. \tag{3.86}$$

□

4 An Analysis of the Adaptation Speed of Causal Models

Prologue to the Second Contribution

Article Details

An Analysis of the Adaptation Speed of Causal Models. *Rémi Le Priol, Reza Babanezhad Harikandeh, Yoshua Bengio, Simon Lacoste-Julien.* This paper was published at AISTATS 2021 (Le Priol et al., 2021a).

Contributions of the Authors

Rémi Le Priol derived the comparison between distances for all models, wrote the code, ran the experiments and contributed to the general writing of the paper. Reza Babenezhad contributed to the general writing and found the optimization method and the convergence rate for normal variables. Simon Lacoste-Julien came up with the idea of using convergence rates to quantify the adaptation speed, while Yoshua Bengio came up with the idea of measuring the adaptation speed in the first place. Simon Lacoste-Julien and Yoshua Bengio supervised this project.

Abstract

Consider a collection of datasets generated by unknown interventions on an unknown structural causal model G. Recently, Bengio et al. (2020) conjectured that among all candidate models, G is the *fastest to adapt* from one dataset to another, along with promising experiments. Indeed, intuitively G has less mechanisms to adapt, but this justification is incomplete. Our contribution is a more thorough analysis of this hypothesis. We investigate the adaptation speed of cause-effect SCMs. Using convergence rates from stochastic optimization, we justify that a relevant proxy for adaptation speed is distance in parameter space after intervention. Applying this proxy to categorical and normal cause-effect models, we show two results. When the intervention is on the cause variable, the SCM with the correct causal direction is advantaged by a large factor. When the intervention is on the effect variable, we characterize the relative adaptation speed. Surprisingly, we find situations where the anticausal model is advantaged,

falsifying the initial hypothesis. Source code for all experiments is hosted at https://github.com/remilepriol/causal-adaptation-speed.

4.1 Introduction

A learning agent interacting with its environment should be able to answer questions such as "what will happen to Y if I change X". Structural Causal Models (SCM) offer a formalism to answer this kind of questions (Pearl, 2009; Peters et al., 2017). The simplest SCM is the model $X \to Y$ where X is the cause and Y the effect. Modifying X will modify Y but modifying Y will not alter X. In general, SCMs model the distribution of observations with a directed graph where edges represent *independent mechanisms* (Janzing and Scholkopf, 2010).

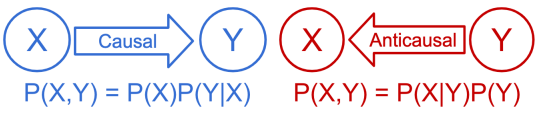


Figure 4.1 – Two models for data (X, Y) with causal structure $X \to Y$.

Modern machine learning methods can fail surprisingly when the test distribution differ from the training distribution (Rosenfeld et al., 2018). A recent line of work describes these distribution shifts as interventions in an underlying causal model (Zhang et al., 2013; Magliacane et al., 2018). If this description is accurate, then an agent endowed with this hypothetical causal model could handle distribution shifts by updating the few mechanisms affected by the intervention. On contrary, an agent endowed with an incorrect model, would have to update many mechanisms. Bengio et al. (2020) infer that the causal agent will be the fastest to adapt to distribution shifts. Conversely, they use the speed of adaptation to unknown interventions as a criterion to learn the true causal model, showing promising empirical results on cause-effect models. Yet they lack a theoretical argument to connect interventions and fast adaptation. Thus we raise the question:

Do causal models adapt faster than non-causal models to distribution shifts induced by interventions?

Contributions. We theoretically and empirically answer this question for cause-effect SCMs with categorical variables, and partially for multivariate normal distributions.

- For both settings, we use stochastic optimization convergence rates to show that the adaptation speed mostly depends on the distance in parameter space between the initialization (before intervention) and the optimum (after intervention).

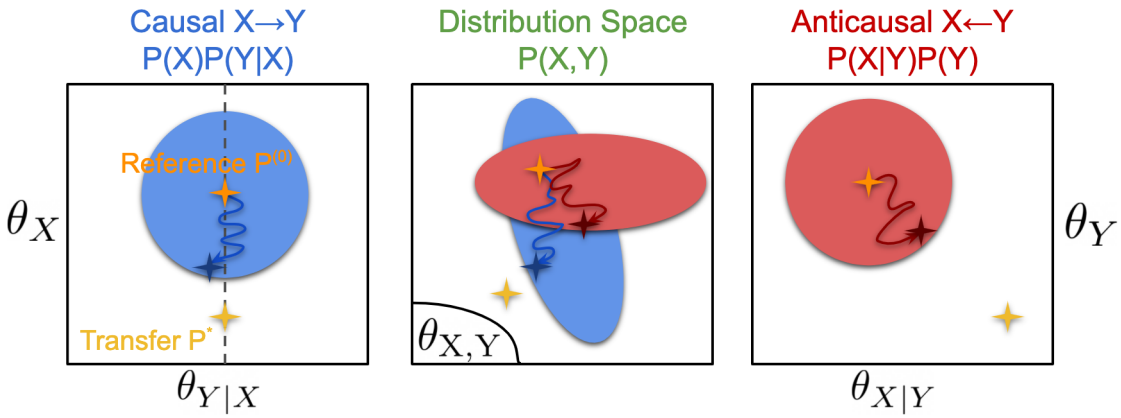


Figure 4.2 – Intuition behind fast adaptation. An intervention on X turns the reference distribution $\boldsymbol{p}^{(0)}$ into a transfer distribution $\boldsymbol{p}^*$. The causal model (blue) only has to adapt θ_X, whereas the anticausal model (red) has to adapt both its mechanisms. After adaptation, the causal model ends up the closest from the transfer in terms of KL, as visible in the abstract distribution space. Blue and red balls represent the *proximity prior* induced by taking a few steps of SGD from the reference in each parameter space. Convergence rate analysis reveals that they are spherical functions of the parameter distance, but they get mapped to non-trivial shapes in distribution space – ellipses in this sketch.

- For categorical variables, we fully characterize this distance. We show that the causal model is faster by a large factor when the intervention is on the cause.
- When the intervention is on the effect, we surprisingly find settings where the anticausal model is systematically faster. As appealing as the fastest-to-adapt hypothesis may sound, it does not hold in every situations.

4.2 Related Work

Causal relationships are asymmetric. These asymmetries are often visible in observations, so that one can identify which is cause and which is effect under relevant assumptions (Mooij et al., 2016). A common assumption is to constrain the set of functional dependencies between cause and effect. By contrast, in our work, we focus on two families of distributions which are notoriously unidentifiable from observational data: categorical and linear normal variables (Peters et al., 2017, Ch.4). With data coming from a generic directed acyclic graph (DAG), we can only hope to discover the Markov equivalence class of this DAG (Verma and Pearl, 1991). Many methods seek to achieve this goal, whether constraint-based such as the PC algorithm (Spirtes et al., 2000) or score-based methods using greedy search

(Chickering, 2002) or more recently continuous optimization (Zheng et al., 2018; Lachapelle et al., 2020). However to discover the exact graph, we need access to interventional data.

Inferring causal links from interventions or experiments is the foundation of science. Inferring causal links from unknown interventions is a much harder and less principled problem. Tian and Pearl (2001) first studied this setting, proposing a constraint based method to infer the interventional equivalence class from a sequence of interventions. Then Eaton and Murphy (2007) proposed an exact Bayesian approach. More recently, Squires et al. (2019); Ke et al. (2019) proposed score based algorithms, improving in scalability and alleviating parametric assumptions. From a machine learning perspective, we are concerned with the predictive power that this structure will give us when faced with new data.

Distribution shifts are a common problem in machine learning, as well as in causal statistics (Zhang et al., 2013; Pearl and Bareinboim, 2014). Schölkopf et al. (2012) first brought up the idea of *invariance* to tackle this problem. Following up on this idea, Peters et al. (2016) designed an algorithm able to identify robust causal features from heterogeneous data. This work has set a fruitful line of research for robust machine learning (Heinze-Deml et al., 2018b,a; Rothenhäusler et al., 2019; Arjovsky et al., 2019). In a way, fast adaptation is the complementary idea of invariance: if most mechanisms are kept invariant, then only a few have to adapt. Schölkopf (2019) shed light on these approaches and the broader scope of causality research for machine learning.

4.3 Background

In this section, we review the formalism of Bengio et al. (2020) on observations, interventions, models and adaptation.

Reference and Transfer Distributions. We assume perfect knowledge of a reference distribution $\boldsymbol{p}$ over the pair (X, Y) sampled from an SCM $X \to Y$. This distribution is the object of interventions, which results in new *transfer* distributions $\boldsymbol{p}^*$. If the *intervention is on the cause*, X is sampled from a different marginal, then Y is sampled from the reference conditional

$$\boldsymbol{p}^*(x, y) = \boldsymbol{p}^*(x)\boldsymbol{p}(y|x) \ . \tag{4.1}$$

If the *intervention is on the effect*, X is sampled from the reference marginal, then Y is sampled from another marginal independently of X

$$\boldsymbol{p}^*(x, y) = \boldsymbol{p}(x)\boldsymbol{p}^*(y) \ . \tag{4.2}$$

For each transfer distribution, we observe a few samples.

Models. We parametrize two generative models of (X, Y) (Fig. 4.1):

$$\boldsymbol{p}_\theta(x, y) = \boldsymbol{p}_{\theta_X}(x)\boldsymbol{p}_{\theta_{Y|X}}(y|x) \quad - \textit{causal} \tag{4.3}$$

$$\boldsymbol{p}_{\theta_\leftarrow}(x, y) = \boldsymbol{p}_{\theta_Y}(y)\boldsymbol{p}_{\theta_{X|Y}}(x|y) \quad - \textit{anticausal}\,. \tag{4.4}$$

For each model, we call mechanisms the marginal and conditional models. Each mechanisms has its own set of parameters, e.g. θ_X and $\theta_{Y|X}$. In the following we will use θ to denote interchangeably $\theta_\rightarrow$ and $\theta_\leftarrow$.

Adaptation. Both models are initialized to fit perfectly the reference distribution $\boldsymbol{p}_{\theta_\rightarrow^{(0)}} = \boldsymbol{p}_{\theta_\leftarrow^{(0)}} = \boldsymbol{p}$. They observe fresh samples from $\boldsymbol{p}^*$ one by one and update their parameters $\theta_\rightarrow$ and $\theta_\leftarrow$ to maximize the log-likelihood with a step of stochastic gradient (SGD). Thanks to the separate parameters, the causal model log-likelihood loss decomposes as

$$\begin{aligned}\mathcal{L}_{\text{causal}}(\theta_\rightarrow) &= \mathbb{E}_{(X,Y)\sim\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_\theta(X, Y)\right] \\ &= \mathbb{E}_{\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_X}(X)\right] + \mathbb{E}_{\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_{Y|X}}(Y|X)\right]\end{aligned} \tag{4.5}$$

When $\boldsymbol{p}^*$ comes from an intervention, Bengio et al. (2020) observe that the causal model is often faster to adapt than the anticausal model. Intuitively, this is because the causal model has to adapt only the mechanism which was modified by the intervention. On the other hand, the anticausal model has to adapt both its mechanisms. In Figure 4.2, we compare these different scenarios and the concept of adaptation figuratively. While appealing, *this reasoning is not rigorous*, as sample complexity bounds of SGD typically do not depend on the number of parameters to update (Bubeck et al., 2015, Th. 6.2 & 6.3). In the next section, we formalize and understand this phenomenon in the light of convergence rates of stochastic optimization methods.

Distribution Families. We study two of the simplest sub-families of the exponential family (Wainwright and Jordan, 2008): categorical and linear normal variables. Their negative log-likelihood is a convex function of their natural parameter. These families are interesting because the direction is not identifiable from observational data (Peters et al., 2017, Ch.4) – e.g. $\boldsymbol{p}_\theta$ and $\boldsymbol{p}_{\theta_\leftarrow}$ can model the same set of distributions – which makes them challenging for causal discovery.

4.4 An Optimization Perspective

One way to formalize adaptation speed is to characterize it via the convergence speed of the stochastic optimization procedure. An appealing aspect of stochastic

optimization algorithms such as SGD (when only using fresh samples and running it on the true loss we care about) is that they come with convergence rate guarantees on the *population risk* in machine learning, thus giving us direct sample complexity results to obtain a specific generalization error. The convergence rate is an *upper bound* on the expected suboptimality after a given number of iterations. While these rates are about worst case performance and might also be loose, fortunately, for convex optimization, they tend to correspond well to actual empirical performance (Nesterov, 2004a). We can thus use the convergence bounds as theoretical proxy for the convergence speed. In our experiments, we also verify empirically that the bounds correlate well with the observed convergence speed.

Here we provide a classical convergence rate on the expected suboptimality with Average Stochastic Gradient Descent (ASGD) under convexity and bounded gradient assumptions. We re-derive this rate in Appendix 4.A.1 for completeness. This rate applies to log-likelihood maximization for categorical random variables (details in 4.A.2). Since the target distribution is part of the model family, the log-likelihood suboptimality is equal to the KL-divergence – e.g. $\mathcal{L}(\theta) - \mathcal{L}(\theta^*) = D_{\mathrm{KL}}(\boldsymbol{p}^* || \boldsymbol{p}_\theta)$.

ASGD. Assume $\forall \theta, x, \|\nabla \log \boldsymbol{p}_\theta(x)\| \leq B$. After T iterations of SGD on (4.5),

$$\theta^{(t+1)} = \theta^{(t)} + \gamma \nabla \log \boldsymbol{p}_{\theta^{(t)}}(X_t, Y_t) \tag{4.6}$$

with learning rate $\gamma := \frac{c}{\sqrt{T}}$, starting from $\theta^{(0)}$, the average parameter's $\bar{\theta}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} \theta^{(t)}$ suboptimality is upper bounded by

$$\mathbb{E}\left[D_{\mathrm{KL}}(\boldsymbol{p}^* || \boldsymbol{p}_{\bar{\theta}^{(T)}})\right] \leq \frac{c^{-1}\|\theta^{(0)} - \theta^*\|^2 + cB^2}{2\sqrt{T}} \tag{4.7}$$

where the expectation is taken over the sampling of $T-1$ training points X_t, Y_t and θ^* is the closest solution to $\theta^{(0)}$ in the solution set $\operatorname{argmin}_\theta \mathcal{L}(\theta)$. For categorical models, $B = 2$ (see 4.A.2). Consequently, for a fixed T and with small enough c, the convergence upper bounds for causal and anticausal models differ mainly by $\delta := \|\theta^{(0)} - \theta^*\|^2$.

The bounded gradient assumption of (4.7) does not apply to the log-likelihood of normal variables. In Section 4.6.1, we provide an algorithm along with a convergence rate (4.22) that do apply to this case. Overall both bounds (4.7) and (4.22) carry the same message which can be summarized by:

The adaptation speed is dominated by
the initial distance

$$\delta_{\mathrm{causal}} = \left\|\theta^{(0)}_{\rightarrow} - \theta^*_{\rightarrow}\right\|^2 \tag{4.8}$$

$$\delta_{\mathrm{anticausal}} = \left\|\theta^{(0)}_{\leftarrow} - \theta^*_{\leftarrow}\right\|^2 . \tag{4.9}$$

Other optimization methods. Yang et al. (2016, Theorem 1) provides a unified convergence rate for stochastic heavy ball and Nesterov methods that is similar to (4.7), where the initial distance is the main difference between causal and anticausal models. Consequently, our theoretical analysis holds for a larger class of algorithms than ASGD. More generally, it applies to any stochastic optimization method whose sample complexity depends on parameter distance.

4.5 Categorical Variables

In this section, both cause and effect come from categorical distribution. We provide theoretical bounds on δ_{causal} and $\delta_{\text{anticausal}}$. We consider different scenarios to generate reference and transfer data and explain the consequences of each scenario.

4.5.1 Definitions

Cause X and effect Y are now two categorical variables taking values in $\{1, \dots, K\}$. Categorical variables are an exponential family with mean parameters $\boldsymbol{p} \in \boldsymbol{\Delta}_K$ the probability vector, and with natural parameter $\boldsymbol{s} \in \mathbb{R}^K$ – the logits or score parameters such that $p_z = \frac{e^{s_z}}{\sum_{z'} e^{s_{z'}}}$. The causal model has parameters $\boldsymbol{s}_X := (s_x)_{x=1\dots K}$ and $\boldsymbol{s}_{Y|X} := (s_{y|x})_{x,y=1\dots K}$. We gather the causal parameters in the variable $\theta_\rightarrow = (\boldsymbol{s}_X, \boldsymbol{s}_{Y|X})$ and the anticausal parameters in $\theta_\leftarrow = (\boldsymbol{s}_Y, \boldsymbol{s}_{X|Y})$ (Fig. 4.3). The loss (4.5) becomes

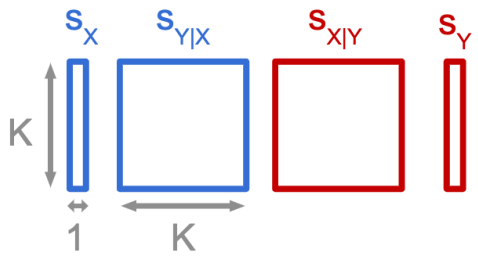


Figure 4.3 – Parametrization of causal (blue) and anticausal (red) categorical models

$$\begin{aligned}\mathcal{L}_{\text{causal}}(\theta_\rightarrow) &= \mathbb{E}_{(X,Y)\sim\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_\theta(X,Y)\right] \\ &= \mathbb{E}_{\boldsymbol{p}^*}\left[-s_X + \log\sum_x e^{s_x} - s_{Y|X} + \log\sum_y e^{s_{y|X}}\right].\end{aligned} \tag{4.10}$$

Each mechanism's stochastic loss is the sum of a linear function and a softmax function. The softmax function is convex and 1-Lipschitz, so we can apply rate (4.7). To be self-contained, we include details in Appendix 4.A.2.

4.5.2 Distance after Intervention

In this section, we prove that interventions on the cause advantage the causal model by a factor K, and we describe when interventions on the effect will advantage

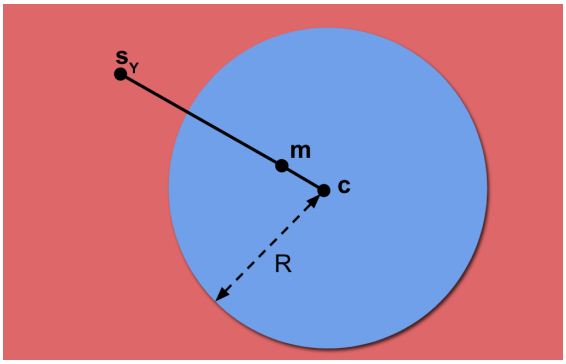


Figure 4.4 – Illustration of Proposition 4.5.2. $\boldsymbol{c}$ is on the line joining $\boldsymbol{s}_Y$ and $\boldsymbol{m} := \frac{1}{K}\sum_x \boldsymbol{s}_{Y|x}$. When $\boldsymbol{s}^*_Y$ is within the blue ball of radius R centered at $\boldsymbol{c}$, $\Delta \leq 0$ and the causal model is advantaged, otherwise the anticausal model is advantaged (red area). This is a surprising counter-example to the adaptation-speed hypothesis.

one model over another.

Intervention on cause X, $\boldsymbol{s}_X \leftarrow \boldsymbol{s}^*_X$. The causal conditional $\boldsymbol{s}_{Y|X}$ is left unchanged, but the effect marginal $\boldsymbol{s}_Y$ is modified in a non-trivial way. Consequently the initial distances are

$$\delta_{\text{causal}} = \|\boldsymbol{s}_X - \boldsymbol{s}^*_X\|^2 \tag{4.11}$$

$$\delta_{\text{anticausal}} = \|\boldsymbol{s}_Y - \boldsymbol{s}^*_Y\|^2 + \sum_y \|\boldsymbol{s}_{X|y} - \boldsymbol{s}^*_{X|y}\|^2 \,. \tag{4.12}$$

The causal model has to update K parameters, whereas the anticausal model has to adapt $K^2 + K$ parameters. Therefore the causal model seems to be advantaged by a factor K. The following proposition – proved in Appendix 4.B – shows that this is reflected by ℓ_2 distances.

Proposition 4.5.1. *When the intervention happens on the cause,*

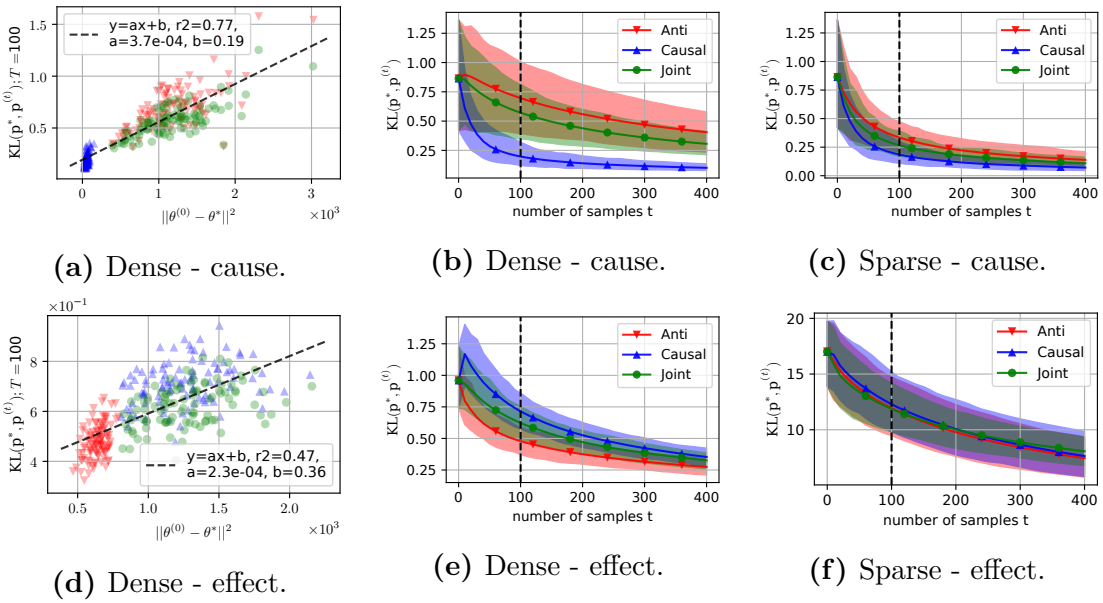
$$\delta_{anticausal} \geq K\delta_{causal} \,. \tag{4.13}$$

Intervention on effect Y, $\forall x, \boldsymbol{s}_{Y|x} \leftarrow \boldsymbol{s}^*_Y$. Cause and effect become independent. The causal model is advantaged only if the intervention $\boldsymbol{s}^*_Y$ is close enough from the previous marginal, as formalized by the following proposition:

Proposition 4.5.2. *When the intervention happens on the effect*

$$\begin{aligned}\Delta &:= \delta_{causal} - \delta_{anticausal} \\ &= (K-1)\left(\|\boldsymbol{s}^*_Y - \boldsymbol{c}\|^2 - R^2\right)\end{aligned} \tag{4.14}$$

where $R^2 \approx K\widehat{\text{Var}}_X[\log \sum_y e^{s_{y|X}}]$ *and* $\boldsymbol{c} = \frac{\left(\sum_x \boldsymbol{s}_{Y|x}\right) - \boldsymbol{s}_Y}{K-1}$.



(a) Dense - cause. **(b)** Dense - cause. **(c)** Sparse - cause.

(d) Dense - effect. **(e)** Dense - effect. **(f)** Sparse - effect.

Figure 4.5 – Experimental results on categorical data. Each plot is captioned with the prior and the intervention considered. **Scatter plots** are showing the positive correlation between the KL after 100 steps of SGD and the initial parameter distance. Each point represent one of 100 synthetic pairs $(\boldsymbol{p}^{(0)}, \boldsymbol{p}^*)$. **Training curves** show the average KL (solid line) and the (5,95) percentiles (shaded) over 100 runs. Remark how all models start from the same initial KL, but they converge at different speeds.

See Figure 4.4 for an illustration and Appendix 4.B.3 for the exact formula of R and the proof. When the intervention $\boldsymbol{s}_Y^*$ is close enough to $\boldsymbol{c}$, which depends on the reference, the causal model is advantaged. If $\boldsymbol{s}_Y^*$ is far from $\boldsymbol{c}$ or if R is small then the anticausal model is likely to be advantaged.

4.5.3 Simulating Reference Distributions

To evaluate the fast adaptation criterion, we are going to work on synthetic data, which raises the question : from which distribution should we sample $\boldsymbol{p} = \boldsymbol{p}_{\theta^{(0)}}$? We call this distribution *prior*. Following the independent mechanism assumption, the marginal on the cause $\boldsymbol{p}_X$ and the conditional of effect given cause $\boldsymbol{p}_{Y|X}$ should not contain any information about each other.

Dense Prior. To sample causal mechanisms, a natural choice is

$$\boldsymbol{p}_X \sim \mathrm{Dir}(\mathbf{1}_K) \quad \text{and} \quad \forall x, \boldsymbol{p}_{Y|x} \sim \mathrm{Dir}(\mathbf{1}_K) \tag{4.15}$$

where Dir is the Dirichlet distribution and $\mathbf{1}_K$ is the all-one vector of dimension K. $\mathrm{Dir}(\mathbf{1}_K)$ the uniform law over the simplex $\boldsymbol{\Delta}_K$. This prior leads to the K2 score

from the Bayesian network literature (Cooper and Herskovits, 1991). We call this choice the *dense prior* by opposition to the sparse prior introduced next. This is the choice made in Bengio et al. (2020), as well as Chalupka et al. (2016). The latter work reports that distributions sampled from this prior exhibit some asymmetry between X and Y. In Appendix 4.C.1, we complement their work, explaining how the effect marginal is likely to be closer from the uniform distribution than the cause marginal. This asymmetry means that *the causal direction is identifiable from observational data.*

Sparse Prior. To fix this issue, we study an alternative prior that is symmetric and ensures that both cause and effect marginals are sampled from a uniform prior over $\boldsymbol{\Delta}_K$. We sample the causal mechanisms as follows

$$\boldsymbol{p}_X \sim \mathrm{Dir}(\mathbf{1}_K) \quad \text{and} \quad \forall x, \boldsymbol{p}_{Y|x} \sim \mathrm{Dir}(\mathbf{1}_K / K) \; . \tag{4.16}$$

The $\mathbf{1}_K / K$ parameter means that samples will be approximately sparse, hence the name. We show in Appendix 4.C.2 that with this sampling scheme, the joint is sampled from a sparse Dirichlet over $\boldsymbol{\Delta}_{K^2}$: $\boldsymbol{p}_{(X,Y)} \sim \mathrm{Dir}(\mathbf{1}_{K^2} / K)$. This in turns means that we can switch the roles of X and Y in (4.16). The effect marginal has uniform density over the simplex. In general, *the causal direction is not identifiable from observational data.* In Bayesian Networks literature, this is known as the Bayesian Dirichlet equivalent uniform prior (Heckerman et al., 1995).

4.5.4 Categorical Variables Experiments

Goal. As discussed in Section 4.5.3, the prior over the joint distribution on (X, Y) is going to influence the behavior of ASGD. We are seeking answers to two questions:

1. Is the adaptation speed positively correlated with the initial distance, as suggested by the upper bound (4.7) on the convergence rate of ASGD?

2. Is there a clear difference in adaptation speed between causal and anticausal models?

Data. We consider categorical variables with $K = 20$. For each initialization method, we sample 100 different reference joint distributions. For each of these distributions, we sample an intervention by sampling a probability vector $\boldsymbol{q}$ *uniformly* from $\boldsymbol{\Delta}_K$. If the intervention is on the cause, we plug $\boldsymbol{q}$ instead of $\boldsymbol{p}_X$. If the intervention is on the effect, we redefine $\boldsymbol{p}_{Y|x} = \boldsymbol{q}, \forall x$.

Models. We are comparing causal and anticausal models adaptation speed. We also report results for a model of the joint $\boldsymbol{p}_{X,Y} = \mathrm{softargmax}(\boldsymbol{s}_{X,Y})$ as a reference model. We expect its results to be in between the performance of the causal and

anticausal model as it expresses no prior over the direction. We optimize all models with Averaged SGD. In each iteration of SGD we get one fresh sample from the transfer distribution. For each model and each setting, we tune the (constant) learning rate so as to optimize the likelihood after seeing $\frac{K^2}{4} = 100$ samples, to explore the few samples regime. We present results in Figure 4.5

Dense prior. When the intervention is on the cause, the causal model is much closer from its optimum: in Fig. 4.5a the blue cluster is on the left of the scatter plots. This is well correlated with faster adaptation (Fig. 4.5b). On the contrary, *when the intervention is on the effect, the anticausal model starts closer from its optimum* and it converges faster (Fig. 4.5d, 4.5e). We can interpret this result in light of Proposition 4.5.2. In Appendix 4.B.3, we explain why the radius R is small under the dense prior. As a result, $\boldsymbol{s}_Y^*$ is mostly sampled outside of the ball of radius R, consequently the anticausal model is advantaged. Overall, there is a wider gap between models in Fig. 4.5b than in Fig. 4.5e. Consequently, if we take a balanced average of a few interventions on the cause and a few interventions on the effect, the causal model remains faster (details in Appendix 4.B.4).

Sparse prior. When the intervention is on the cause, the causal model has a slight advantage (Fig. 4.5c). When the intervention is on the effect, no model has a set advantage (Fig. 4.5f), but the sparsity induces much higher KL values, as explained in Appendix 4.C.3. This KL explosion drowns the signal coming from the cause intervention, calling for further algorithmic developments – such as inferring the intervention, as explored by Ke et al. (2019).

4.6 Multivariate Normal Variables

In this section, we analyze the case of two multivariate normal variables with a linear relationship. Cause X and effect Y are sampled from the causal model

$$X \sim \mathcal{N}(\mu_X, \Sigma_X) \tag{4.17}$$
$$Y|X \sim \mathcal{N}(\boldsymbol{A}X + \boldsymbol{a}, \Sigma_{Y|X}) \tag{4.18}$$

with mean parameters $\mu_X, \boldsymbol{a} \in \mathbb{R}^K$ and $\Sigma_X, \boldsymbol{A}, \Sigma_{Y|X} \in \mathbb{R}^{K\times K}$. This parametrization is the most intuitive but it is unfortunately not appropriate to get convergence rates. We are going to introduce another parametrization along with an algorithm and a convergence rate (Sec. 4.6.1), before providing empirical results (Sec. 4.6.2).

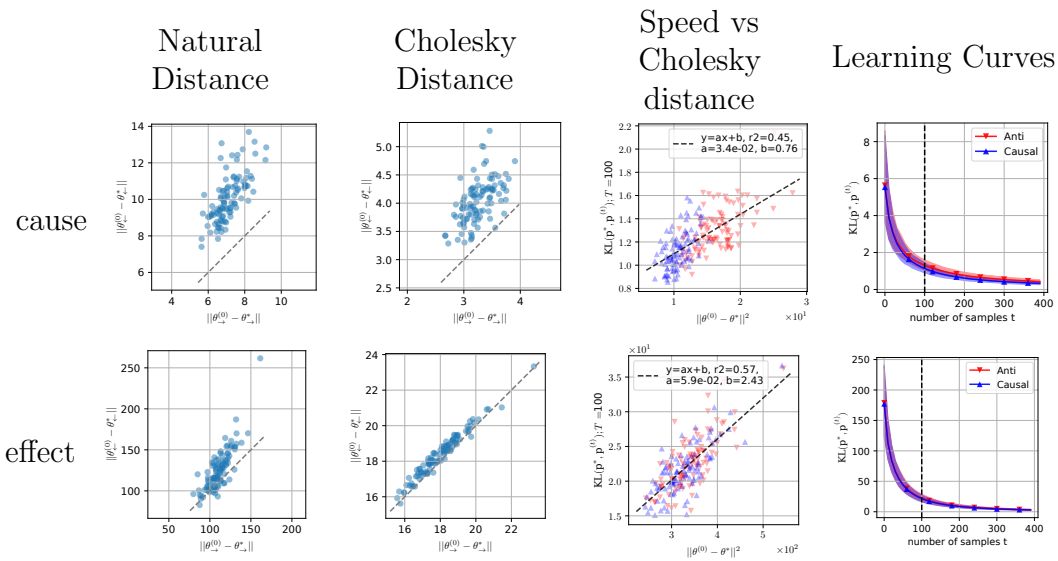


Figure 4.6 – Multivariate Normal Variables with dimension $K = 10$. Row 1 and 2 correspond to interventions on cause and effect respectively. *Column 1 & 2:* scatter plot $\delta_{\text{anticausal}}$ vs δ_{causal} respectively in natural and Cholesky parametrization. The grey diagonal is the identity line. We observe a natural tendency for $\delta_{\text{anticausal}} > \delta_{\text{causal}}$ (points above the grey diagonal), but this is systematically true only for the natural distance when the intervention is on the cause. *Column 3 & 4:* same plot as in Figure 4.5. Once again we observe a correlation between initial distance and optimization speed. When the intervention is on the cause, the causal model is advantaged. When the intervention is on the effect, both curves overlap.

4.6.1 Optimization Analysis

The negative log-likelihood of model (4.17) is notoriously non-convex. This is problematic for convergence results. For simplicity, we focus in this section on the simple marginal mechanism with mean parameters μ, Σ. We detail the full model in Appendix 4.E. If we use the natural parameters $\eta = \Sigma^{-1}\mu$ and $\Lambda = \Sigma^{-1}$ (precision matrix), the negative log-likelihood is convex

$$\begin{aligned} &\mathbb{E}\left[-\log \boldsymbol{p}_{(\eta,\Lambda)}(X)\right] \\ &= \frac{1}{2}\Big(\mathbb{E}\left[\mathrm{Tr}(XX^\top\Lambda) - 2X^\top\eta\right] + \eta^\top\Lambda^{-1}\eta - \log|\Lambda|\Big)\,. \end{aligned} \tag{4.19}$$

This objective is composed of a pleasant stochastic linear term, and a difficult deterministic barrier objective which goes to infinity when $\Lambda \to 0$. This barrier is composed of a matrix inverse and a log determinant. The assumptions of Lipschitz or gradient-Lipschitz required to get SGD convergence do not hold for the barrier. While the empirical version of (4.19) has a close formed formula for its global minimum, quite surprisingly, gradient-based optimization of the normal likelihood is difficult to analyze. Convex optimization typically deals with non-smooth terms by

introducing proximal operators (Parikh et al., 2014). However this barrier term is too complex to get an analytic formula for the proximal operator. We transform it into a more convenient form by introducing $\boldsymbol{L}$, the lower triangular Cholesky factor of the precision matrix $\Lambda = \boldsymbol{L}\boldsymbol{L}^T$, and $\zeta = \boldsymbol{L}^{-1}\eta = \boldsymbol{L}^\top \mu$. Then (4.19) simplifies into

$$\begin{aligned} &\mathbb{E}\left[-\log \boldsymbol{p}_{(\zeta,\boldsymbol{L})}(X)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left\|\boldsymbol{L}^\top X - \zeta\right\|^2\right] - \sum_i \log \boldsymbol{L}_{i,i} \ . \end{aligned} \tag{4.20}$$

We will refer to $(\zeta, \boldsymbol{L})$ as *Cholesky parameters*. This objective is more suitable to gradient based optimization with a simple proximal operator, as detailed in the next section. We provide all details about the causal model in Appendix 4.E.

Stochastic Proximal Gradient Algorithm We want to minimize the sum of a stochastic convex smooth function $f_X(\theta) := \frac{1}{2}\left\|\boldsymbol{L}^\top X - \zeta\right\|^2$ and convex non-smooth regularizer $g(\theta) = -\sum_i \log \boldsymbol{L}_{i,i}$. This is exactly the goal of the stochastic proximal gradient (Duchi et al., 2010) update

$$\theta_{t+1} = \operatorname*{argmin}_\theta g(\theta) + \frac{1}{2\gamma_t}\|\theta_t - \gamma_t \nabla f_{X_t}(\theta_t) - \theta\|^2 \tag{4.21}$$

where γ_t is the step-size and X_t is randomly sampled. For objective (4.20), the proximal gradient update has a closed form solution that amounts to updating all parameters with the stochastic gradient of the quadratic term, then updating the diagonal elements of $\boldsymbol{L}$ with the mapping $x \mapsto \frac{1}{2}(x + \sqrt{x^2 + 4\gamma})$, thus ensuring that they remain strictly positive (details in Appendix 4.D.2).

Convergence Rate. We assume that stochastic gradients are almost-surely B-Lipschitz. B is known as the smoothness constant. We show in Appendix 4.D.1 that running the stochastic proximal gradient algorithm with step size $\gamma_t = \frac{\gamma}{3B\sqrt{T}}$ where $\gamma \leq 1$, for T iterations guarantees

$$\mathbb{E}\left[D_{\mathrm{KL}}(\boldsymbol{p}^* || \boldsymbol{p}_{\bar{\theta}^{(T)}})\right] \leq \frac{3B\|\theta^{(0)} - \theta^*\|^2}{\gamma\sqrt{T}} + \frac{D_{\mathrm{KL}}(p^* || p_{\theta^{(0)}})}{T} \ . \tag{4.22}$$

Analysis. The term $KL(p^*||p_{\theta^{(0)}})/T$ is equal for causal and anticausal models because we assume $\boldsymbol{p}_\theta^{(0)} = \boldsymbol{p}_{\theta_\leftarrow}^{(0)}$. For normal variables, B depends only on the data and is a priori equal for both models (Appendix 4.D.3). Similarly to (4.7), both models' rates differ mainly by $\delta = \|\theta^{(0)} - \theta^*\|^2$.

When the intervention is on the cause, we prove in Appendix 4.E.5 that the anticausal model is farther away from its optimum in the natural parametrization

$$\delta_{\text{anticausal}}^{\text{natural}} \geq \delta_{\text{causal}}^{\text{natural}} \ . \tag{4.23}$$

Unfortunately, in the Cholesky parametrization (Fig. 4.6, 2nd column), or when the intervention is on the effect (Fig. 4.6, bottom row),we observe empirically that there is no such hard guarantee, although the causal distance tends to be smaller than the anticausal distance.

4.6.2 Experiments

Similarly to categorical variables, we need to decide on a prior over reference and transfer distributions. This choice is informed by two criteria. First the independent mechanism principle which states that we should sample θ_X independently of $\theta_{Y|X}$. Second we want θ_Y to have approximately the same distribution as θ_X – e.g. we want the distribution to be approximately symmetric so that we cannot identify the direction from observational data. These considerations lead us to a flavor of normal-Wishart prior (Geiger et al., 2002) described in Appendix 4.F.

We sample 100 random joint distributions from this prior, and for each distribution we sample a random intervention on the cause, and a random intervention on the effect. We then run the stochastic proximal gradient on objective (4.20). We report results in Figure 4.6. Similarly to the categorical case, when the intervention is on the cause, the causal model is advantaged by a slight margin (upper right figure). When the intervention is on the effect both models are learning at the same speed (bottom right figure).

Conclusion

We provided a first theoretical analysis of the adaptation speed in two-variables cause-effect SCMs under localized interventions for categorical and normal data. Convergence guarantees for stochastic optimization on the true population log-likelihood indicates that the adaptation speed is related to the distance between initial point and optimum in parameter space. We verified this correlation empirically. We proved analytically that this distance is lower for the causal model than for the anticausal model when the intervention is on the cause variable. This explains a surprising phenomenon: while both models start with the same suboptimality, one learns faster than the other. When the intervention is on the effect variable, we highlighted examples showing that either model can be advantaged. This observation challenges the intuition that the causal model should be the fastest to adapt, and it raises new questions for the approach of Bengio et al. (2020), such as: are there practical situations where the fastest-to-adapt heuristic is useful ? On a more theoretical note, is it possible to characterize the adaptation speed behavior for more general families of distributions?

4.A Categorical Optimization

In this section we prove a convergence rate of ASGD that applies to the categorical loss, and we show that the constants involved in this rate are the same for both causal and anticausal models.

4.A.1 Convergence of ASGD with Fixed Step-Size

Here we derive a classical convergence rate of Average SGD. This result is standard ; we include it to be self-contained. The objective is

$$\min_{\theta} F(\theta) = \mathbb{E}_i \left[f(\theta, i) \right] \ . \tag{4.24}$$

Theorem 4.A.1. *If each $f_i(\theta) = f(\theta, i)$ has bounded gradient B, then after T steps of SGD with step-size $\gamma = \frac{c}{\sqrt{T}}$, starting from θ_0, the expected sub-optimality verifies*

$$\mathbb{E}\left[F(\bar{\theta}_T) - F(\theta^*) \right] \leq \frac{1}{2c\sqrt{T}} \|\theta_0 - \theta^*\|^2 + \frac{cB^2}{2\sqrt{T}} \tag{4.25}$$

where $\bar{\theta}_T = \frac{1}{T} \sum_t \theta_t$.

Proof. First we relate the ℓ^2 distance to optimum at step $t+1$ with the one at step t :

$$\begin{aligned}
&\|\theta_{t+1} - \theta^*\|^2 \\
&= \|\theta_t - \theta^*\|^2 - 2\gamma \left\langle f_i'(\theta_t), \theta_t - \theta^* \right\rangle + \gamma^2 \|f_i'(\theta_t)\|^2 \\
&\leq \|\theta_t - \theta^*\|^2 - 2\gamma \left\langle f_i'(\theta_t), \theta_t - \theta^* \right\rangle + \gamma^2 B^2 \ .
\end{aligned}$$

By convexity of f_i and rearranging the terms we get

$$\begin{aligned}
2\gamma(f_i(\theta_t) - f_i(\theta^*)) &\leq 2\gamma \left\langle f_i'(\theta_t), \theta_t - \theta^* \right\rangle \\
&\leq \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2 + \gamma^2 B^2.
\end{aligned}$$

Now we take the expectation, sum up both sides for T iterations and divide by $2T\gamma$ to get

$$\begin{aligned}
&\frac{1}{T} \sum_{i=1}^{T} \mathbb{E}\left[F(\theta_t) - F(\theta^*) \right] \\
&\leq \frac{1}{2\gamma T} \left(\mathbb{E}\left[\|\theta_0 - \theta^*\|^2 \right] - \mathbb{E}\left[\|\theta_{T+1} - \theta^*\|^2 \right] \right) + \frac{\gamma B^2}{2} \\
&\leq \frac{1}{2\gamma T} \|\theta_0 - \theta^*\|^2 + \frac{\gamma B^2}{2}
\end{aligned}$$

Finally, we apply Jensen inequality to F in $\bar{\theta}_T = \frac{1}{T} \sum_t \theta_t$ to get the final result. □

4.A.2 Categorical Loss Properties

We are now going to verify that assumptions of the rate (4.7) apply to the negative log-likelihood loss for the categorical distribution. This loss is standard and it's properties are well-known, but we review them here to be self-contained.

Each mechanism has the same form of negative log-likelihood, with the same kind of stochastic gradients. The total loss is a sum over mechanisms, and the total stochastic gradient is a concatenation of each mechanisms stochastic gradient. To apply rate 4.7, we can either apply it separately on each mechanism, either apply to the whole. Both path lead to the same result. In the end, we simply have to check that this loss is convex, and has bounded gradients for all z. The random functions coming from sampling Z are

$$f_z(\boldsymbol{s}) = -s_z + \log(\sum_{z'} e^{s_{z'}}) \tag{4.26}$$

This function is the softmax – or logsumexp – function minus a stochastic linear term.

Convexity We are going to show that it is convex but not strongly convex because it becomes flat for large score values. Its derivative is

$$\nabla f_z(\boldsymbol{s}) = -\boldsymbol{e}_z + \boldsymbol{p}\ . \tag{4.27}$$

where $\boldsymbol{e}_z$ is the z-th canonical basis element and $\boldsymbol{p}$ is the output of the softargmax function taken on $\boldsymbol{s}_Z$. The Hessian is the same for every z.

$$\nabla^2 f_z(\boldsymbol{s}) = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\ . \tag{4.28}$$

We observe that for any vector $\boldsymbol{v}$,

$$\begin{aligned}\boldsymbol{v}^\top \nabla^2 f_z(\boldsymbol{s})\boldsymbol{v} &= \sum_z p_z v_z^2 - \left(\sum_z p_z v_z\right)^2 \\ &= \operatorname{Var}_{Z\sim\boldsymbol{p}}[v_Z] \geq 0\end{aligned} \tag{4.29}$$

which means that the logsumexp is convex. When s_0 tends toward positive infinity and the other components remain constant, $\boldsymbol{p}$ tends toward a Dirac on the 0-th component. Then (4.29) is 0 for all $\boldsymbol{v}$, so the logsumexp is not strongly convex.

Bounded Gradients. The gradient norm is

$$\|\nabla f_z(\boldsymbol{s})\| = \|\boldsymbol{p} - \boldsymbol{e}_z\| \tag{4.30}$$

$$. \tag{4.31}$$

This norm is maximized for $\boldsymbol{p} = \boldsymbol{e}_{z'}, \forall z' \neq z$. The maximum is equal to $\sqrt{2}$. If there are d independent mechanisms (for d variables in the graph), then the total stochastic gradient which is a concatenation of all gradients has a norm bounded by $B = \sqrt{2d}$. In our case of cause-effect models, $d = 2$ and and the gradients are bounded by $B = 2$, or in other words, all the f_z are 2-Lipschitz.

This bound is the same for causal and anticausal models. It depends on the part of space where $\boldsymbol{p}$ is going to live. Assuming that it is going to live in most of the space for both directed models, both loss will have the same Lipschitz constants in practice.

Thanks to these properties, the sample complexity of $\boldsymbol{p}_\theta$ and $\boldsymbol{p}_{\theta_\leftarrow}$ are bounded by (4.7). The difference in adaptation speed between causal and anticausal models is characterized by the distance in parameter space.

4.B Categorical Analysis

In this section, we prove relationships between parameter distances induced by interventions between the causal and anticausal models. First we prove two useful lemmas. Then we establish that the causal model dominates the anticausal model by a factor K when the intervention is on the cause. Finally we show that no model has a set advantage when the intervention bears on the effect.

The logits or scores $\boldsymbol{s}$ live in $\mathbb{R}^K$. They have one additional degree of freedom compared to the probability $\boldsymbol{p}$. More specifically, the softargmax is invariant by translations along the vector $\mathbf{1} = (1, \dots, 1)$. In other words, all scores $\{\boldsymbol{s} + \lambda\, \mathbf{1} \,|\forall \lambda \in \mathbb{R}\}$ are equivalent. Scores which move by following the gradient of this loss will remain in the same affine hyperplane orthogonal to $\mathbf{1}$. To ensure that the distances we measure are meaningful, we project all logits in the hyperplane such that $\sum_z s_z = 0$, by subtracting their mean.

Definition 4.B.1 (Mean-zero score)**.** *A score vector $\boldsymbol{s}$ is mean-zero iff $\sum_z s_z = 0$.*

4.B.1 Switching Direction

In this section we are going to prove a few useful results relating cause and anticausal models. We know the causal parameters $X \to Y$, and we want to find the corresponding $X \leftarrow Y$ model, e.g. express $\boldsymbol{s}_Y, \boldsymbol{s}_{X|Y}$ as a function of $\boldsymbol{s}_X, \boldsymbol{s}_{Y|X}$. This will help us to find a relationship between δ_{causal} and $\delta_{\text{anticausal}}$. We first need to define a few useful variables

Definition 4.B.2 (Average conditional score vectors)**.** *For any x or y, define*

$$m(y) := \frac{1}{K} \sum_x s_{y|x}, \quad n(x) := \frac{1}{K} \sum_y s_{x|y} \,. \tag{4.32}$$

Definition 4.B.3 (Conditional log-partition function)**.**

$$A(x) = \log \sum_y e^{s_{y|x}} \tag{4.33}$$

With these variables, we can express the reverse conditional score from the causal parameters.

Lemma 4.B.4 (Anticausal conditional score)**.** *Let s_y, s_x be marginal scores, and $s_{y|x}$, $s_{x|y}$ be conditional scores. Then*

$$\boxed{s_{x|y} = s_x + (s_{y|x} - m(y)) - (A(x) - \alpha)}, \tag{4.34}$$

where $\alpha = \frac{1}{K}\sum_x A(x)$.

Proof. Let's apply Bayes rule to find the conditional probability mass function

$$\begin{aligned} \boldsymbol{p}(x|y) &\propto \boldsymbol{p}(y|x)\boldsymbol{p}(x) \\ &\propto \exp\left(s_{y|x} - A(x) + s_x\right) \end{aligned}$$

where $A(x)$ is the log-partition function of $\boldsymbol{p}(y|x)$. Taking the logarithm,

$$s_{x|y} = s_x + s_{y|x} - A(x) + C(y) \tag{4.35}$$

where $C(y)$ is a constant defined such that $\sum_x \boldsymbol{s}_{x|y} = 0$ (see Definition 4.B.1). We take the sum of (4.35) over x to find

$$\begin{aligned} & \sum_x s_x + & \sum_x s_{y|x} - & \sum_x A(x) + & KC(y) \\ = \quad & 0 + & Km(y) + & K\alpha + & KC(y) \end{aligned}$$

which simplifies into

$$C(y) = -m(y) - \alpha$$

We plug this in (4.35) to conclude the proof. □

We conclude this section with an identity showing that conditional logits are equally close from their averages in both directions.

Lemma 4.B.5. *For any x and y we have*

$$\boxed{s_{x|y} - n(x) = s_{y|x} - m(y)}. \tag{4.36}$$

where $n(x) := \frac{1}{K}\sum_y s_{x|y}$ and $m(y) := \frac{1}{K}\sum_x s_{y|x}$.

Proof. We apply Lemma 4.B.4 with the roles of X and Y inverted to express $\boldsymbol{s}_{y|x}$ as a function of the anti causal parameters

$$s_{y|x} = s_y + (s_{x|y} - n(x)) - (B(y) - \beta) \,, \tag{4.37}$$

where $B(y) = \log \sum_x e^{s_{y|x}}$ and $\beta = \frac{1}{K} \sum_y B(y)$. We can add (4.34) and (4.37) to get rid of the conditional scores

$$s_x - n(x) - A(x) + \alpha = -(s_y - m(y) - B(y) + \beta),$$

for all x, y. The left hand side is constant in y whereas the right hand side is constant in x. Thus both sides are constants with respect to both x and y. In particular they are equal to their average

$$\begin{aligned} \forall x, s_x - n(x) - A(x) + \alpha &= \frac{1}{K} \sum_{x'} (s_{x'} - n(x')) \\ &\quad - \frac{1}{K} \sum_{x'} A(x') + \alpha \\ &= 0 - 0 - \alpha + \alpha \\ &= 0 \,. \end{aligned}$$

We plug this equality into (4.34) to prove the lemma. □

4.B.2 Intervention on Cause

In this section, we analyze the relationship between δ_{causal} and $\delta_{\text{anticausal}}$ after an intervention on the cause.

Proposition 4.5.1. *If an intervention happens on the cause X then we have*

$$\boxed{\delta_{anticausal} \geq K \delta_{causal}} \,, \tag{4.38}$$

where $\delta_{causal} = \|\boldsymbol{s}_X - \boldsymbol{s}^*_X\|^2$, and $\delta_{anticausal} = \|\boldsymbol{s}_Y - \boldsymbol{s}^*_Y\|^2 + \sum_y \|\boldsymbol{s}_{X|y} - \boldsymbol{s}^*_{X|y}\|^2$

Proof. Given that $s^*_{y|x} = s_{y|x}, \boldsymbol{m}^* = \boldsymbol{m}, A^* = A$ and $\alpha^* = \alpha$, Lemma 4.B.4 tells us that the anticausal conditional $\boldsymbol{s}^*_{X|Y}$ verifies

$$\begin{aligned} s^*_{x|y} - s^*_x = (s_{y|x} - m(y)) - (A(x) - \alpha) = s_{x|y} - s_x \\ \implies s_{x|y} - s^*_{x|y} = s_x - s^*_x \,. \end{aligned}$$

The distance between models before and after intervention are

$$\begin{aligned} \delta_{\text{causal}} &= \|\boldsymbol{s}_X - \boldsymbol{s}^*_X\|^2 \\ \delta_{\text{anticausal}} &= \|\boldsymbol{s}_Y - \boldsymbol{s}^*_Y\|^2 + \sum_y \|\boldsymbol{s}_{X|y} - \boldsymbol{s}^*_{X|y}\|^2 \\ &\geq 0 + \sum_y \sum_x (s_x - s^*_x)^2 = K \|\boldsymbol{s}_X - \boldsymbol{s}^*_X\|^2 \,. \end{aligned}$$

In conclusion,

$$\delta_{\text{anticausal}} \geq K\delta_{\text{causal}} \ . \tag{4.39}$$

□

4.B.3 Intervention on Effect

The following proposition shows that when the intervention is on the effect, the causal model is advantaged only when the new effect marginal $\boldsymbol{s}_Y^*$ is close enough from the previous marginal.

Proposition 4.5.2 *When an intervention happens on the effect*

$$\Delta := \delta_{causal} - \delta_{anticausal} \tag{4.40}$$

$$= (K-1)\left(\|\boldsymbol{s}_Y^* - \boldsymbol{c}\|^2 - R^2\right) \tag{4.41}$$

where the score vector $\boldsymbol{c}$ *and the scalar* R *are defined as*

$$\boldsymbol{c} = \frac{K\boldsymbol{m} - \boldsymbol{s}_Y}{K-1} \tag{4.42}$$

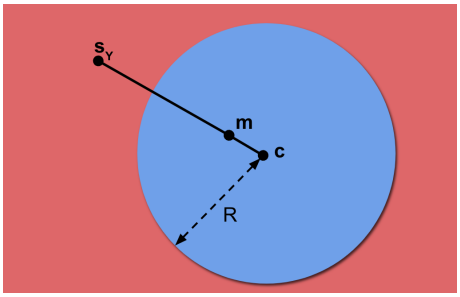
$$\begin{aligned}(K-1)R^2 =& K\|\boldsymbol{n} - \boldsymbol{s}_X\|^2 + (K-1)\|\boldsymbol{c}\|^2 \\ &+ \|\boldsymbol{s}_Y\|^2 - K\|\boldsymbol{m}\|^2\end{aligned} \tag{4.43}$$

with $\boldsymbol{m}$ *and* $\boldsymbol{n}$ *as in Definition 4.B.2.*

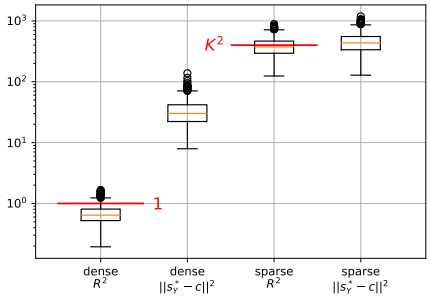
We illustrate the relationship between $\boldsymbol{m}$, $\boldsymbol{c}$, $\boldsymbol{s}_Y$ and R in Figure 4.7a.

Proof. First we expand the causal distance with a bias variance decomposition

$$\begin{aligned}\delta_{\text{causal}} &= \sum_x \|\boldsymbol{s}_{Y|x} - \boldsymbol{s}_Y^*\|^2 \\ &= \sum_{x,y}(s_{y|x} - m(y) + m(y) - s_y^*)^2 \\ &= \sum_{x,y}(s_{y|x} - m(y))^2 + K\sum_y (m(y) - s_y^*)^2 \\ &\quad + 2\sum_y (m(y) - s_y^*)\underbrace{\sum_x (s_{y|x} - m(y))}_{=0} \ .\end{aligned} \tag{4.44}$$



(a) $\boldsymbol{m}$ is a convex combination of $\boldsymbol{c}$ and $\boldsymbol{s}_Y$. The blue bubble is the sub-level set 0 of Δ. It is a sphere of radius R centered at $\boldsymbol{c}$. Within this sphere, $\delta_{\text{causal}} \leq \delta_{\text{anticausal}}$ the causal model is advantaged. Outside this sphere, $\delta_{\text{causal}} \geq \delta_{\text{anticausal}}$ the anticausal model is advantaged.



(b) Box plots for the radius R^2 and deviations $\|\boldsymbol{s}_Y^* - \boldsymbol{c}\|^2$ for $K = 20$ with the dense prior (left) and the sparse prior (right). The y-axis is logarithmic. Red lines show analytical estimates for the expected radius.

Figure 4.7 – Schematic and numerical illustrations of Proposition 4.5.2.

Given that $\boldsymbol{s}^*_{X|y} = \boldsymbol{s}_X$, we can decompose the anticausal distance similarly

$$\begin{aligned}
\delta_{\text{anticausal}} &= \|\boldsymbol{s}_Y - \boldsymbol{s}_Y^*\|^2 + \sum_y \|\boldsymbol{s}_{X|y} - \boldsymbol{s}^*_{X|y}\|^2 \\
&= \|\boldsymbol{s}_Y - \boldsymbol{s}_Y^*\|^2 + \sum_{x,y} (s_{x|y} - n(x) + n(x) - s_x)^2 \\
&= \|\boldsymbol{s}_Y - \boldsymbol{s}_Y^*\|^2 + \sum_{x,y} (s_{x|y} - n(x))^2 \\
&\quad + K \sum_x (n(x) - s_x)^2 \,. \qquad (4.45)
\end{aligned}$$

Thanks to Lemma 4.B.5, the variance of conditional score vectors (in blue) in (4.44) and (4.45) are equal

$$\sum_{x,y} (s_{x|y} - n(x))^2 = \sum_{x,y} (s_{y|x} - m(y))^2 \,.$$

What remains in the difference is the quadratic form

$$\begin{aligned}
\Delta &= \delta_{\text{causal}} - \delta_{\text{anticausal}} \\
&= K\|\boldsymbol{m} - \boldsymbol{s}_Y^*\|^2 - K\|\boldsymbol{n} - \boldsymbol{s}_X\|^2 - \|\boldsymbol{s}_Y - \boldsymbol{s}_Y^*\|^2 \,,
\end{aligned}$$

which we can expand to highlight the role of $\boldsymbol{s}_Y^*$ as

$$\begin{aligned}\Delta =&(K-1)\|\boldsymbol{s}_Y^*\|^2 - 2\langle \boldsymbol{s}_Y^*, K\boldsymbol{m} - \boldsymbol{s}_Y\rangle \\ &+ K\|\boldsymbol{m}\|^2 - \|\boldsymbol{s}_Y\|^2 - K\|\boldsymbol{n} - \boldsymbol{s}_X\|^2\|^2 \\ =&(K-1)\|\boldsymbol{s}_Y^* - \boldsymbol{c}\|^2 - (K-1)\|\boldsymbol{c}\|^2 \\ &+ K\|\boldsymbol{m}\|^2 - \|\boldsymbol{s}_Y\|^2 - K\|\boldsymbol{n} - \boldsymbol{s}_X\|^2\end{aligned}$$

where $\boldsymbol{c}$ appears as a non-convex interpolation of $\boldsymbol{m}$ and $\boldsymbol{s}_Y$, $\boldsymbol{c} = \frac{K\boldsymbol{m} - \boldsymbol{s}_Y}{K-1}$. Define R^2 to conclude the proof. □

Empirical estimates of the radius. We report values of R^2 and $\|\boldsymbol{s}_Y^* - \boldsymbol{c}\|^2$ observed for the dense and sparse priors in Figure 4.7b. For dense prior radii are much smaller than deviations, whereas for the sparse prior they have similar magnitude. This explains why the anticausal model systematically adapts faster when the intervention is on the effect and the prior is dense. We also observe that radii (and deviations) are much greater for the sparse prior than for the dense prior. In the following paragraph we provide some clues to explain this behaviour.

As illustrated by Figure 4.7a, $\boldsymbol{m}$ is a convex combination of $\boldsymbol{c}$ and $\boldsymbol{s}_Y$: $\boldsymbol{m} = \frac{(K-1)\boldsymbol{c} + \boldsymbol{s}_Y}{K}$ so by convexity of $\|.\|^2$,

$$\begin{aligned}&\frac{K-1}{K}\|\boldsymbol{c}\|^2 + \frac{1}{K}\|\boldsymbol{s}_Y\|^2 \geq \|\boldsymbol{m}\|^2 \\ \implies &(K-1)\|\boldsymbol{c}\|^2 + \|\boldsymbol{s}_Y\|^2 - K\|\boldsymbol{m}\|^2 \geq 0 \\ \implies &R^2 \geq \|\boldsymbol{n} - \boldsymbol{s}_X\|^2 \ .\end{aligned}$$

As K grows larger, this inequality will get closer and closer to an equality. Indeed, $\boldsymbol{m}$ will get closer and closer to $\boldsymbol{c}$ and we will end up with

$$\frac{K-1}{K}\|\boldsymbol{c}\|^2 + \frac{1}{K}\|\boldsymbol{s}_Y\|^2 - \|\boldsymbol{m}\|^2 \ll \|\boldsymbol{n} - \boldsymbol{s}_X\|^2 \approx R^2 \ .$$

Before proceeding, let us prove a simple proposition that is a direct consequence of Lemma 4.B.4.

Proposition 4.B.6. *The squared distance between marginal score and reverse average conditional is equal to the empirical variance of the conditional log-partition function*

$$\|\boldsymbol{s}_X - \boldsymbol{n}\|^2 = K\widehat{\mathrm{Var}}_X[A(X)] \ . \tag{4.46}$$

Proof. Taking the average of Lemma 4.B.4 over y yields

$$\begin{aligned}\frac{1}{K}\sum_y s_{x|y} &= s_x - (A(x) - \alpha) + \frac{1}{K}\sum_y (s_{y|x} - m(y)) \\ n(x) &= s_x - (A(x) - \alpha) \ ,\end{aligned}$$

where we used the definition of $n(x)$ and the mean-zero scores. Reordering terms gives

$$s_x - n(x) = A(x) - \alpha \tag{4.47}$$

Recall that α is the average of $A(x)$ over x. Squaring this equation and summing over x concludes the proof. □

Using this proposition we get that

$$\begin{aligned} R^2 &\approx \|\boldsymbol{s}_X - \boldsymbol{n}\|^2 \\ &= K\widehat{\mathrm{Var}}_X[A(X)] \\ &= K\widehat{\mathrm{Var}}_X[\log \sum_y e^{s_{y|X}}] \, . \end{aligned}$$

Conditional scores $\boldsymbol{s}_{y|x}$ are taking much greater values with much higher variance under the sparse prior than under the dense prior. To be clear, we sample independently pseudo-scores $\tilde{s}_{y|x}$ from exp-gamma laws, and we subtract their mean to ensure that they sum to 0 (Definition 4.B.1). $s_{y|x} = \tilde{s}_{y|x} - \frac{1}{K}\sum_{y'} \tilde{s}_{y'|x}$. This means that

$$\log \sum_y e^{s_{y|X}} = \log \sum_y e^{\tilde{s}_{y|X}} - \frac{1}{K}\sum_y \tilde{s}_{y|x}$$

If we make the approximation that the logsumexp term and the average term are independent then

$$\begin{aligned} &\widehat{\mathrm{Var}}_X[\log \sum_y e^{s_{y|X}}] \\ &\approx \mathrm{Var}_{s_{y|x}}[\log \sum_y e^{s_{y|x}}] \\ &\approx \mathrm{Var}_{\tilde{s}_{y|x}}[\log \sum_y e^{\tilde{s}_{y|X}}] + \frac{1}{K}\mathrm{Var}_{\tilde{s}_{y|x}}[\sum_y \tilde{s}_{y|x}] \\ &= \psi^{(1)}(K\lambda) + \frac{1}{K}\psi^{(1)}(\lambda) \end{aligned}$$

where this last step uses the formula for the variance of an exp-gamma variable twice. The variance of an exponential gamma with shape parameter λ is $\psi^{(1)}(\lambda)$ where $\psi^{(1)}$ is the trigamma function. The log-sum-exp of K independent exp-gamma with scale and shape parameters (λ, ζ) is another exp-gamma with scale and shape parameters $(K\lambda, \zeta)$. Finally we get the following approximation for the squared radius

$$R^2 \approx K\psi^{(1)}(K\lambda) + \psi^{(1)}(\lambda) \, .$$

The dense prior uses a shape parameter $\lambda = 1$ while the sparse prior uses a shape parameter $\lambda = \frac{1}{K}$. We use two approximations of the trigamma function: $\psi^{(1)}(\frac{1}{K}) \approx K^2 + \pi^2/6$ and $\psi^{(1)}(K) \approx \frac{1}{K}$ when $K \geq 10$.

$$\begin{aligned} R^2_{\text{dense}} &\approx K\psi^{(1)}(K) + \psi^{(1)}(1) \\ &\approx 1 + \pi^2/6 = O(1) \\ R^2_{\text{sparse}} &\approx K\psi^{(1)}(1) + \psi^{(1)}(\frac{1}{K}) \\ &\approx K^2 + K\pi^2/6 + \pi^2/6 = O(K^2) \ . \end{aligned}$$

In other words for dense prior the radius grows linearly with the dimension K. We report these estimates along with real data in Figure 4.7b.

Independent Special Case. if X is independent of Y in the reference distribution – e.g. $\forall x, y, \boldsymbol{p}(x, y) = \boldsymbol{p}(x)\boldsymbol{p}(y)$ – then $\forall x, y$,

$$\begin{aligned} s_x &= s_{x|y} = n(x) \\ s_y &= s_{y|x} = m(y) \end{aligned}$$

Plugging these equalities into Proposition 4.5.2 yields

$$\begin{aligned} \boldsymbol{c} = \boldsymbol{m} = \boldsymbol{s}_Y \quad \text{and} \quad R = 0 \\ \implies \Delta = (K-1)\|\boldsymbol{s}^*_Y - \boldsymbol{s}_Y\|^2 \geq 0 \end{aligned}$$

which means that the anticausal model is advantaged $\delta_{\text{causal}} \geq \delta_{\text{anticausal}}$. This is actually predictable from a simple parameter counting argument. When the reference distribution is made of independent distributions, the anticausal conditional mechanism is already optimal $s_x = s_{x|y}$. The anticausal model only has to adapt its marginal mechanism $\boldsymbol{s}_Y$ of size K. On contrary, the causal model only has to adapt its conditional mechanism $s_{y|x} \neq s_y$ of size K^2. Overall the causal model has to adapt K times more parameters than the anticausal model.

4.B.4 Other Empirical Results for Cause and Effect Interventions

In this section, we present additional results for categorical variables. In Figure 4.8, compared to the main text, we add what happens with the dense prior when we average learning curves (pooled) from 5 interventions on the cause and 5 interventions on the effect : on average the causal model adapts the fastest. In Figure 4.9, compared to the main text we show what happens with the sparse prior, both in terms of distance (scatter plots) and in terms of pooled results. Because the intervention on the effect creates huge values of the KL, there is no set advantage for any of the models.

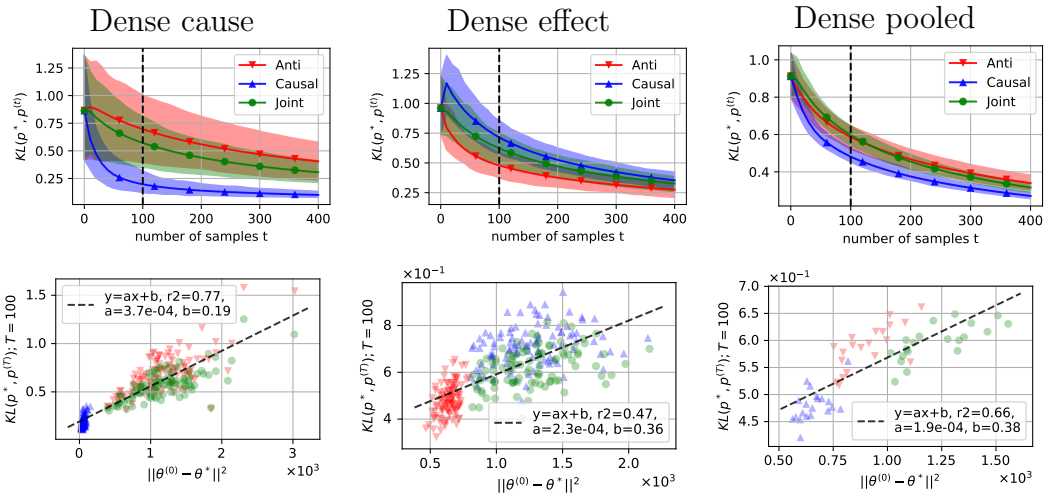


Figure 4.8 – Categorical dense prior with K=20. *Row 1:* training curves. Solid lines are average KL over 100 runs. We tune hyper-parameters to minimize the average KL of each model at the black vertical dashed bar (t=100). Shaded areas are between (5,95) quantiles. Note that *all models start from the same initial KL, but they converge at different speeds. Row 2:* scatter plot of the KL at t=100 vs. initial distance. Note that the initial distance is well correlated with the KL after 100 steps of SGD. *Columns:* we report results for interventions on the cause on column 1, the effect on column 2, and an aggregation of both on column 3. We aggregate results by taking the average of 5 cause interventions and 5 effect interventions as one new trajectory. In total we have 20 such trajectories per model. We are reporting this result because the meta-learning criterion suggested by Bengio et al. (2020) is akin to the average adaptation speed over a small set of interventions.

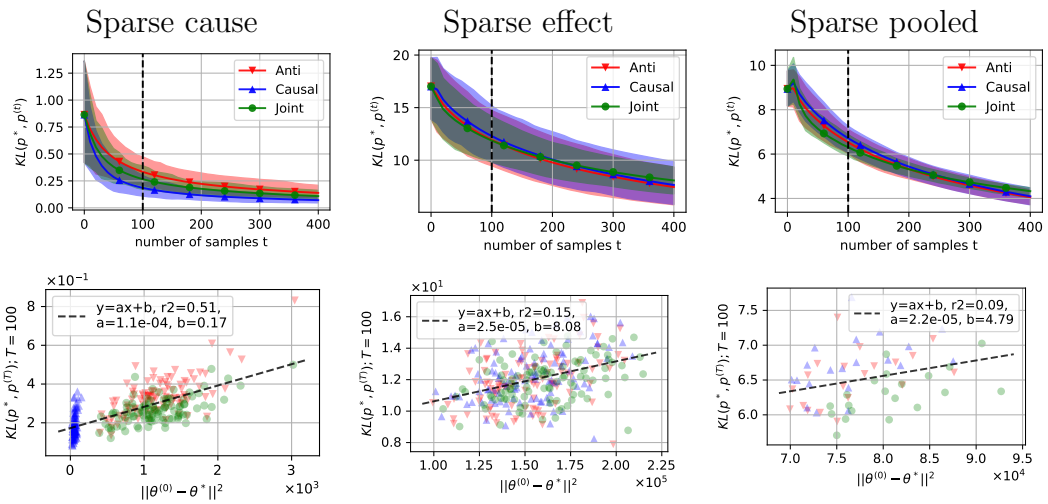


Figure 4.9 – Categorical sparse prior with K=20. *Column 1:* intervention on the cause. The causal model starts closer from optimum and adapts slightly faster than others. *Column 2:* intervention on the effect. All models have the same initial distance and the same objective value. However the KL value is around 10. This is 10 times larger than when the intervention is on the cause. *Column 3:* we take the average of 5 effect and 5 cause interventions. The effect dominates this average because it is much larger. As a result there is no signal.

4.B.5 Single Mechanism Intervention

If only $\boldsymbol{s}_{Y|x_0}$ changes, for some x_0, then from Lemma 4.B.4 we get the following equality

$$\begin{aligned}\delta_{\text{anticausal}} =& \frac{K-1}{K}\delta_{\text{causal}} + \left\|\boldsymbol{s}_Y^* - \boldsymbol{s}_Y\right\|^2 \\ &+ (K-1)(A^*(x_0) - A(x_0))^2 \,. \end{aligned} \tag{4.48}$$

The causal and anticausal distances seem to be on the same scale, with a multiplicative factor $\frac{K-1}{K} \lessapprox 1$ and a positive additive factor. This is interesting because the sparsity argument holds: the causal model needs to change K parameters whereas the anticausal model needs to change $K^2 + K$ parameters. That means we could expect an advantage by a factor K for the causal model, similarly to when the intervention is on the cause. However (4.48) tells another story: without further assumptions, it seems like both distances will have the same scale.

Experiments. For this kind of intervention to be detectable, we need to intervene on x_0 such that $\boldsymbol{p}(x_0)$ is quite large. To ensure this in our experiments, we pick $x_0 = \operatorname{argmax}_x \boldsymbol{p}(x)$. We report results on dense and sparse priors in Figure 4.10. We observe no significant advantage for the causal model, in spite of the parameter counting prediction.

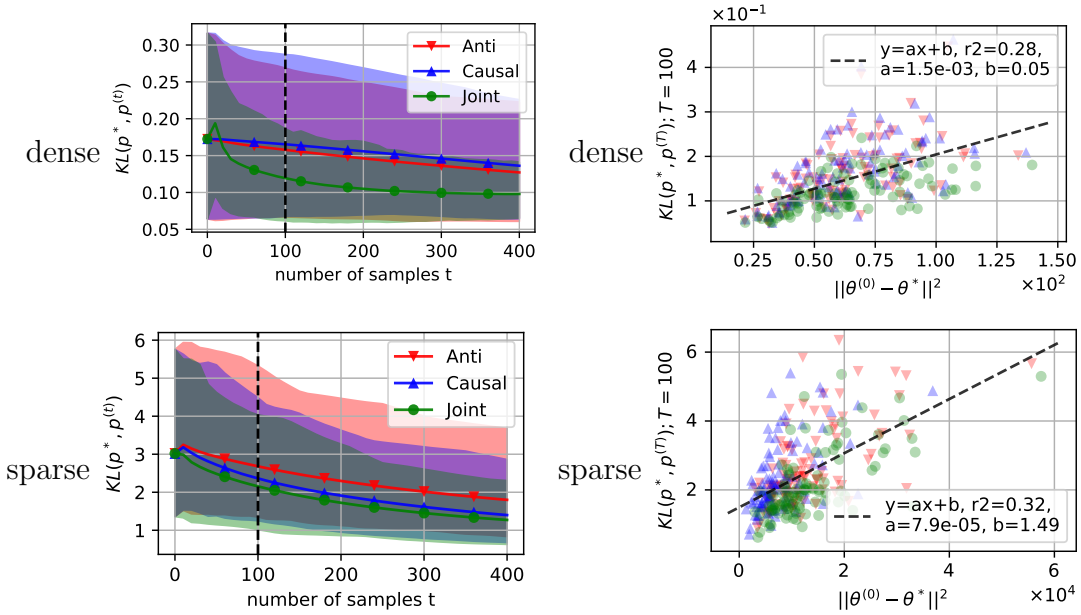


Figure 4.10 – Single mechanism intervention with K=20. *Two first rows:* Dense prior. The only model to be slightly advantaged is the joint model. *Two last rows:* Sparse prior. This time there is a slight advantage for the causal model which performs comparably to the joint model. Overall the optimization is hard in both settings, since we are observing only K^2 samples for models with $O(K^2)$ parameters. The KL barely decreases.

K	2	3	4	5	6	7	8	9	10	11	12	13	14
error	.4	.2	.1	.07	.03	.01	.005	.002	.0007	.0003	7e-5	3e-5	4e-6

Table 4.1 – Estimation of the Bayes error under the dense prior assumption for increasing categorical variables dimension K. We estimated these numbers by sampling one million joint distributions for each K. We report 1 significant figure.

4.C Categorical Priors

In this section we study the dense and sparse prior described in the main paper.

4.C.1 Causal Direction is Identifiable under the Dense Prior

Chalupka et al. (2016) study the prediction of causal direction from observational data under the dense prior assumption. The causal direction $X \to Y$ induces a certain prior over joint distributions $\pi(\boldsymbol{p}|\to)$. The anticausal direction $X \leftarrow Y$ induces another one $\pi(\boldsymbol{p}|\leftarrow)$. The Bayes classifier is predicting $\to$ if

$$\log \pi(\to|\boldsymbol{p}) - \log \pi(\leftarrow|\boldsymbol{p}) > 0 \,, \tag{4.49}$$

and $\leftarrow$ otherwise. Under the dense prior assumption, this classifier makes an error of approximately 0.4 for $K = 2$, which decreases exponentially to 0.001 for $K = 10$. We reproduced their setting and report the error of the optimal classifier in Table 4.1 for varying K. In other words the dense prior induce very asymmetric distributions which makes the causal direction identifiable.

Is this Bayes classifier easy to estimate ? It turns out that under the dense prior, the criterion (4.49) can be simplified into the following criterion

$$D_{\mathrm{KL}}(\boldsymbol{u}||\boldsymbol{p}_X) - D_{\mathrm{KL}}(\boldsymbol{u}||\boldsymbol{p}_Y) > 0 \tag{4.50}$$

where $\boldsymbol{u} := \mathbf{1}/K$ is the uniform probability vector. The proof is left as an exercize to the reader. If (4.50) is positive, Bayes predicts that the cause is X, otherwise Y is the cause. In words, whichever variable has the most uniform marginal is the effect. This simple rule is optimal given the prior assumption (and if both directions are equally likely). We can understand it from a concentration of measure perspective. The effect marginal is written as a sum of quasi independent uniform variables

$$\boldsymbol{p}(y) = \sum_x \boldsymbol{p}(y|x)\boldsymbol{p}(x) \tag{4.51}$$

which ends up close from the uniform vector.

4.C.2 Joint Distribution with Sparse Prior

The following theorem shows how a Dirichlet prior over joint distributions $c_{x,y} = \boldsymbol{p}(x, y)$ is equal to independent Dirichlet priors over marginal $a_x = \boldsymbol{p}(x)$ and conditional $b_{y|x} = \boldsymbol{p}(y|x)$ probability mass functions. By applying this theorem, we find that the sparse prior is equivalent to $\mathrm{Dir}(\frac{1}{K}\mathbf{1}_{K^2})$.

Theorem 4.C.1 (Dirichlet and Factorization)**.** *Let $\boldsymbol{c}$ be a random square matrix of dimension K. Let's define $\boldsymbol{a}$ as the random vector obtained by summing columns of $\boldsymbol{c}$, and $\boldsymbol{b}$ as a copy of $\boldsymbol{c}$ with rows normalized so that they sum to 1. $\forall (i,j) \in \{1,\dots,K\}^2$*

$$a_i = \sum_j c_{i,j} \tag{4.52}$$

$$b_{j|i} = \frac{c_{i,j}}{a_i} \ . \tag{4.53}$$

Let γ be a positive square matrix of parameters. The following equivalence holds

$$\boldsymbol{c} \sim \mathrm{Dir}(\gamma) \iff \begin{cases} \boldsymbol{a} \sim \mathrm{Dir}(\sum_i \gamma_i) \\ \boldsymbol{b}_{:|i} \sim \mathrm{Dir}(\gamma_{i,:}), \forall i \\ \boldsymbol{a} \perp\!\!\!\perp \boldsymbol{b}_{:|i} \perp\!\!\!\perp \boldsymbol{b}_{:|i'}, \forall i \neq i' \end{cases} . \tag{4.54}$$

Proof. First let's remark that the right side of (4.54) is entirely characterizing the joint distribution on $(\boldsymbol{a}, \boldsymbol{b})$, and that the relationship between $\boldsymbol{c}$ and $(\boldsymbol{a}, \boldsymbol{b})$ is a bijection with reverse $c_{i,j} = b_{j|i}a_i$. This means that the equivalence (4.54) is an equality between distributions. This means that we can prove the forward implication and the converse will hold automatically.

If $\boldsymbol{c} \sim \mathrm{Dir}(\gamma)$, then there exist K^2 independent Gamma variables $\tilde{c}_{i,j} \sim \Gamma(\gamma_{i,j}, 1), \forall i, j$ such that

$$\boldsymbol{c} = \frac{\tilde{\boldsymbol{c}}}{S} \quad \text{where} \quad S = \sum_{i,j} \tilde{c}_{i,j} \ . \tag{4.55}$$

We know from properties of the Gamma distribution that $\boldsymbol{c}$ is independent of S. Now let's define $\tilde{a}_i := \sum_j \tilde{c}_{i,j}$. This definition has three consequences. First $\tilde{a}_i$ is a sum of independent gammas, so it is a gamma with parameters $(\alpha_i := \sum_j \gamma_{i,j}, 1)$. Second $S = \sum_i \tilde{a}_i$. Third $\boldsymbol{a} = \frac{\tilde{\boldsymbol{a}}}{S}$ is independent of S and is a Dirichlet with parameter vector $\boldsymbol{\alpha} = \sum_j \boldsymbol{\gamma}_{:,j}$.

That was for the marginal. Now for the conditional,

$$b_{j|i} = \frac{c_{i,j}}{a_i} = \frac{\tilde{c}_{i,j}/S}{\tilde{a}_i/S} = \frac{\tilde{c}_{i,j}}{\tilde{a}_i} \ . \tag{4.56}$$

Again from properties of the Gamma distribution, $\boldsymbol{b}_{:|i} \perp\!\!\!\perp \tilde{a}_i, \forall i$, and $\boldsymbol{b}_{:|i} \sim \mathrm{Dir}(\gamma_{i,:})$. Each of the conditional $\boldsymbol{b}_{:|i}$ is defined with independent gammas, so we also have the independence between conditionals. We verified all properties of the right side of (4.54), which concludes the proof. □

4.C.3 Categorical Sparse Prior Explosion

In this section we explain why the KL takes large values with sparse prior and effect intervention.

On one hand the sparse prior samples probability vectors which are close from being Dirac. On the other hand the effect intervention creates an outer product between two samples drawn uniformly from the simplex $\mathrm{Dir}(\mathbf{1})$.

For instance, for the uniform probability vector $\boldsymbol{u} = \frac{1}{K^2}\mathbf{1} \in \boldsymbol{\Delta}_{K^2}$ and an almost Dirac $\boldsymbol{p} = (1-\varepsilon)\boldsymbol{e}_1 + \varepsilon\boldsymbol{u}$

$$D_{\mathrm{KL}}(\boldsymbol{u}||\boldsymbol{p}) \in \Theta(\log(\frac{1}{\varepsilon}))$$

where ε is a small value. As we increase K, the sparse prior $\mathrm{Dir}(\mathbf{1}_{K^2}/K)$ becomes more sparse. Conceptually, the value of ε decreases, and the value of $D_{\mathrm{KL}}(\boldsymbol{u}||\boldsymbol{p})$ explodes. This is why we observe high KL values for sparse prior and effect intervention. Empirically, these values also increase with K.

4.D Normal Optimization

In this Section we adapt the stochastic composite mirror-prox algorithm to our setting of unbounded multivariate normal optimization. First we describe the algorithm and prove a novel convergence rate that applies to our setting. Then we explicit the update formulas for the normal log-likelihood loss with Cholesky parameters. Finally we prove that worst case constants appearing in the rate are equal for both causal and anticausal models.

4.D.1 Stochastic Composite Mirror-Prox

We want to minimize the composite objective

$$F(\theta) = \mathbb{E}_i\left[f(\theta, i)\right] + g(\theta)\ .$$

For simplicity we denote $f(\theta, i)$ by $f_i(\theta)$ and $f(\theta) = \mathbb{E}_i\left[f_i(\theta)\right]$. We assume that f_i is convex, ∇f_i is L-Lipschitz and g is a convex function. The stochastic mirror-prox algorithm update rule at time t is

$$\nu_t = \theta_{t+1} - \gamma_t f_i'(\theta_t) \tag{4.57}$$

$$\theta_{t+1} = \operatorname*{argmin}_{\theta}\left\{g(\theta) + \frac{1}{\gamma_t}\mathcal{B}_h(\theta, \nu_t)\right\} \tag{4.58}$$

where f_i is sampled randomly. $B_h(x, y) = h(x) - h(y) - < h'(y), x - y >$ denotes the Bregman divergence between x and y induced by the convex function h and we have

$\|B_h(x,y)\| \geq \frac{\alpha}{2}\|x-y\|^2$. When we set $h(x) = 1/2\|x\|^2$, we recover something called the proximal stochastic gradient method (Duchi and Singer, 2009), also known as Perturbed proximal gradient algorithm (Atchadé et al., 2017). This last citation in particular has hypothesis very close to ours.

Convergence Rate

The following Theorem is a mild modification of the Theorem 8 in (Duchi et al., 2010). Our result is different in 2 ways. First, we remove the boundedness assumption for the Bregman divergence throughout the trajectory i.e. $\mathcal{B}_h(\theta^*, \theta_t) \leq D$ for all t. Second, we replace the f_i B-Lipschitz continuous assumption by ∇f_i B-Lipschitz continuous. We need this last modification for the result to hold on f_i quadratic.

First we prove the following lemma which is a modification of Lemma 1 in (Duchi et al., 2010).

Lemma 4.D.1. *With f convex and B-smooth, g convex, and $\gamma \leq \frac{\alpha}{3B}$, at iteration t, if we sample i, we have:*

$$\begin{aligned}\gamma\left\{f_i(\theta_t) + g(\theta_{t+1}) - F(\theta^*)\right\} \leq & \mathcal{B}_h(\theta^*, \theta_t) - \mathcal{B}_h(\theta^*, \theta_{t+1}) \\ & + \gamma\left\{f_i(\theta_t) - f_i(\theta_{t+1})\right\}.\end{aligned}$$

Proof. We have the following sequence of inequality

$$\begin{aligned}&\gamma\Big(f_i(\theta_t) + g(\theta_{t+1}) - F(\theta^*)\Big) \\ &\leq \gamma\left\langle \theta_t - \theta^*, f_i'(\theta_t)\right\rangle + \gamma\left\langle \theta_{t+1} - \theta^*, \partial g_t(\theta_t)\right\rangle \\ &\leq \mathcal{B}_h(\theta^*, \theta_t) - \mathcal{B}_h(\theta^*, \theta_{t+1}) \\ &\quad - \mathcal{B}_h(\theta_{t+1}, \theta_t) + \gamma\left\langle \theta_t - \theta_{t+1}, f_i'(\theta_t)\right\rangle \\ &\leq \mathcal{B}_h(\theta^*, \theta_t) - \mathcal{B}_h(\theta^*, \theta_{t+1}) \\ &\quad - \mathcal{B}_h(\theta_{t+1}, \theta_t) + \frac{B\gamma}{2}\|\theta_t - \theta_{t+1}\|^2 \\ &\quad + \gamma\Big(f_i(\theta_t) - f_i(\theta_{t+1})\Big)\end{aligned}$$

where the first inequality comes from convexity of f_i and g,the second inequality comes from Eq.(6) of lemma 1 in (Duchi et al., 2010), and the third inequality comes from the smoothness of f_i. By $\gamma \leq \frac{\alpha}{3L}$, the term $-\mathcal{B}_h(\theta_{t+1}, \theta_t) + \frac{B\gamma}{2}\|\theta_t - \theta_{t+1}\|^2$ is negative : we can drop this term and get the required result. Note that ∂g is a subgradient of g. □

Theorem 4.D.2. *Given the above assumptions for f_i and g, after T iterations of the stochastic mirror prox algorithm with $\gamma = \frac{c}{\sqrt{T}}, (c \leq \frac{\alpha}{3B})$, we have*

$$\mathbb{E}\left[F(\bar{\theta}) - F(\theta^*)\right] \leq \frac{\mathcal{B}_h(\theta^*, \theta_0)}{c\sqrt{T}} + \frac{(F(\theta_0) - F(\theta^*))}{T}.$$

Proof. The proof is similar to the proof of the Theorem 8 in (Duchi et al., 2010) with some modifications. Take the expectation of lemma 4.D.1 with respect to the samples $(i_u)_{u\leq t}$

$$\begin{aligned}
&\mathbb{E}\left[\gamma\Big(f(\theta_t)+g(\theta_{t+1})-F(\theta^*)\Big)\right]\\
&\leq \mathbb{E}\left[\mathcal{B}_h(\theta^*,\theta_t)-\mathcal{B}_h(\theta^*,\theta_{t+1})+\gamma\Big(f(\theta_t)-f(\theta_{t+1})\Big)\right]
\end{aligned}$$

where θ_t and θ_{t+1} are random variable that depends on the samples. Sum up both side for T iterations:

$$\begin{aligned}
&\gamma\sum_{t=0}^{T}\mathbb{E}\left[f(\theta_t)+g(\theta_{t+1})-F(\theta^*)\right]\\
&\leq \mathcal{B}_h(\theta^*,\theta_0)-\mathbb{E}\left[\mathcal{B}_h(\theta^*,\theta_{T+1})\right]\\
&+\gamma\Big(f(\theta_0)-\mathbb{E}\left[f(\theta_{T+1})\right]\Big)
\end{aligned}$$

By adding $\gamma\mathbb{E}\left[g(\theta_0)-g(\theta_{t+1})\right]$ to both sides of the above inequality we get:

$$\begin{aligned}
&\gamma\sum_{t=1}^{T}\mathbb{E}\left[F(\theta_t)-F(\theta^*)\right]\\
&\leq \mathcal{B}_h(\theta^*,\theta_0)-\mathbb{E}\left[\mathcal{B}_h(\theta^*,\theta_{T+1})\right]\\
&+\gamma\Big(F(\theta_0)-\mathbb{E}\left[F(\theta_{T+1})\right]\Big)\\
&\leq \mathcal{B}_h(\theta^*,\theta_0)+\gamma\Big(F(\theta_0)-\mathbb{E}\left[F(\theta^*)\right]\Big)
\end{aligned}$$

where the last inequality is due to non-negativity of Bregman divergence and optimality of θ^*. Divide both sides by $\gamma T = c\sqrt{T}$ and use Jensen inequality on F to conclude the proof. □

4.D.2 Normal Model Updates

The objective function at hand is:

$$\begin{aligned}
F(L,\zeta) &= f(L,\zeta)+g(L)\\
f(L,\zeta) &= \frac{1}{2n}\sum_{i=1}^{n}\|L^Tx_i-\zeta\|^2\\
g(L) &= -\ln(|L|).
\end{aligned}$$

Now the update rule for the ζ given that g is independent of ζ and we sample mini-batch B of size m:

$$\zeta_{t+1} = (1-\gamma)\zeta_t+\gamma L_t^T\Big(\frac{1}{m}\sum_{i\in B}x_i\Big)$$

For the L the gradient update gives

$$L_{t+\frac{1}{2}} = L_t - \gamma \frac{1}{m} \sum_{i \in B} (x_i x_i^T L_t - x_i \zeta_t^T) \ .$$

Since L is lower triangular, $g(L) = \ln(|L|) = \sum_{i=1}^{d} \log L_{i,i}$ and the proximal operator only applies to diagonal elements of L – e.g. when $i \neq j$ $[L_{t+1}]_{(i,j)} = [L_{t+\frac{1}{2}}]_{(i,j)}$. Otherwise we have to compute:

$$[L_{t+1}]_{(i,i)} = \underset{L_{ii}}{\operatorname{argmin}} \Big\{ -\ln(L_{ii}) + \frac{1}{2\gamma} \|L_{ii} - [L_{t+\frac{1}{2}}]_{(i,i)}\|^2 \Big\}.$$

Therefore the update rule for the $[L_{t+1}]_{(i,i)}$ is:

$$[L_{t+1}]_{(i,i)} = \frac{1}{2} \Big\{ [L_{t+\frac{1}{2}}]_{(i,i)} + \sqrt{[L_{t+\frac{1}{2}}]^2_{(i,i)} + 4\gamma} \Big\} \ .$$

Remark how this proximal operator behaves as a smooth projection on the set of strictly positive numbers. If the diagonal is negative after the gradient update, it brings it to a small positive value. If it was already positive, it slightly increases its value.

4.D.3 Equality of Smoothness Constants

In this section, we show that the Lipschitz smoothness parameter B, which appears in the convergence rate (4.22), is the same for both causal and anticausal models. Similarly to the categorical case, we reason about marginals loss first because they have a simpler form.

The loss of a marginal mechanism is $f_x(L, \zeta) = \frac{1}{2} \sum_{i=1}^{d} (\zeta_i - L_i^T x)^2$ where x is a sample observation and the L_i are the columns of L. We need to show that its Hessian is upper-bounded $\|\nabla^2 f_x(L, \zeta)\| \leq B$. Thanks to the objective f_x being quadratic, the Hessian is independent of the parameters (L, ζ). It depends only on the data x. Since the data domain is *a priori* the same for causal and anticausal models – e.g. X and Y can live in the same range – the upper bound for the Hessian is the same. This holds true at least for marginal mechanisms, because their loss is written exactly like above.

For conditional mechanisms, this is a bit more complicated but the reasoning holds. The objective is similar, with extra parameters coming from the linear relationship between X and Y. The Cholesky parametrization is described in equation (4.68). The conditional model uses $\zeta_{Y|X} = MX + m$ where M is a matrix, m is a vector and X is a given sample. This means that the objective is still a quadratic and that 2nd order derivatives w.r.t. $L_{Y|X}$ are still independent of the parameters M and m. They depend only on the observed values of X and Y. We assume that these variables have the same domain *a priori*, therefore both models have similar worst case smoothness constants.

4.E Normal Analysis

In this section we introduce three different parametrization of the multivariate normal cause-effect model. The mean parametrization is the most common and intuitive, but it yields a non-convex optimization problem. The natural parametrization yields a convex problem with convergence guarantees, but it has no closed update formulas for our optimization algorithm of choice. The Cholesky parametrization offers both a convex problem and simple updates.

Then we proceed to study how interventions induce distance in parameter space. We prove that in the natural parameter space, an intervention on the cause will create more distance in the anticausal model than in the causal model.

4.E.1 Mean Parameters

Cause X and effect Y are sampled from the causal model

$$\begin{aligned} X &\sim \mathcal{N}(\mu_X, \Sigma_X) \\ Y|X &\sim \mathcal{N}(AX + a, \Sigma_{Y|X}) \end{aligned}$$

with parameters $(\mu_X, \Sigma_X, A, a, \Sigma_{Y|X})$. All along, we will assume that all normal laws are non-degenerate – e.g. $\Sigma_X > 0, \Sigma_{Y|X} > 0$. We compute the marginal mean and covariance of Y as well as the covariance between X and Y

$$\begin{aligned} \mathbb{E}[Y] &= A\mu_X + a \\ \mathrm{Cov}[Y] &= \Sigma_{Y|X} + A\Sigma_X A^\top \\ \mathrm{Cov}[X, Y] &= \Sigma_X A^\top \ . \end{aligned}$$

From there we can derive the joint distribution as a function of the causal parameters

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ A\mu_X + a \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_X A^\top \\ A\Sigma_X & \Sigma_{Y|X} + A\Sigma_X A^\top \end{pmatrix} \right) \ . \tag{4.59}$$

4.E.2 Natural Parameters

We want the negative log-likelihood objective to be convex, so we are going to use the natural parameters instead of the mean parameters

$$\mathcal{N}(\mu, \Sigma) = \mathcal{N}_{\mathrm{nat}}(\eta = \Sigma^{-1}\mu, \Lambda = \Sigma^{-1}) \tag{4.60}$$

where we are using $\mathcal{N}_{\mathrm{nat}}$ to explicit that this is taking the natural parameters as arguments Our causal model using natural parameters is:

$$\begin{aligned} X &\sim \mathcal{N}_{\mathrm{nat}}(\eta_X, \Lambda_X) \\ Y|X &\sim \mathcal{N}_{\mathrm{nat}}(BX + b, \Lambda_{Y|X}) \end{aligned} \tag{4.61}$$

where we get the natural parameters from the mean parameters with formulas

$$\begin{aligned}
\Lambda_X &= \Sigma_X^{-1} \\
\Lambda_{Y|X} &= \Sigma_{Y|X}^{-1} \\
\eta_X &= \Lambda_X \mu_X \\
B &= \Lambda_{Y|X} A \\
b &= \Lambda_{Y|X} a
\end{aligned}$$

Switching Direction

We want to get the natural parameters of the anticausal model as a function of the causal parameters. To do so we are going to express the natural parameters of the joint, and then we will simply have to swap rows and columns to invert the roles of X and Y. To get the joint precision matrix, we need to invert the joint covariance. We use the Schur complement and the blockwise matrix inversion formulas

$$\begin{aligned}
&M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \\
&M/D := D - CA^{-1}B \\
&M^{-1} = \\
&\begin{pmatrix} A^{-1} + A^{-1}B(M/D)^{-1}CA^{-1} & -A^{-1}B(M/D)^{-1} \\ (M/D)^{-1}CA^{-1} & (M/D)^{-1} \end{pmatrix}
\end{aligned}$$

In our case, the Schur complement of Σ with respect to its lower right block Σ_Y is precisely

$$\begin{aligned}
M/D &= \Sigma/\Sigma_Y \\
&= \Sigma_{Y|X} + A\Sigma_X A^\top - A\Sigma_X \Sigma_X^{-1} \Sigma_X A^\top \\
&= \Sigma_{Y|X} \ .
\end{aligned}$$

By applying the formula and identifying the natural parameters, we get

$$\Lambda = \begin{pmatrix} \Lambda_X + B^\top \Lambda_{Y|X}^{-1} B & -B^\top \\ -B & \Lambda_{Y|X} \end{pmatrix}$$

To get the first natural parameter, all we have to do is to multiply the joint precision and the joint mean

$$\begin{aligned}
\eta &= \Lambda\mu \\
&= \begin{pmatrix} \Lambda_X \mu_X + B^\top \Lambda_{Y|X}^{-1} B\mu_X - B^\top A\mu_X - B^\top a \\ -B\mu_X + \Lambda_{Y|X} A\mu_X + \Lambda_{Y|X} a \end{pmatrix} \\
&= \begin{pmatrix} \eta_X - B^\top \Lambda_{Y|X}^{-1} b \\ b \end{pmatrix}
\end{aligned}$$

where $B^\top \Lambda_{Y|X}^{-1} B\mu_X - B^\top A\mu_X$, and $-B\mu_X + \Lambda_{Y|X} A\mu_X$ are zero and we express the other terms with natural parameters. Overall, the joint natural parameters are

$$\mathcal{N}_{\text{nat}}\left(\begin{pmatrix}\eta_X - B^\top \Lambda_{Y|X}^{-1} b \\ b\end{pmatrix}, \begin{pmatrix}\Lambda_X + B^\top \Lambda_{Y|X}^{-1} B & -B^\top \\ -B & \Lambda_{Y|X}\end{pmatrix}\right) .$$

From there we can use a symmetry argument to switch from causal $X \to Y$ to anticausal $X \leftarrow Y$ model.

$$\begin{aligned} Y &\sim \mathcal{N}_{\text{nat}}(\eta_Y, \Lambda_Y) \\ X|Y &\sim \mathcal{N}_{\text{nat}}(CY + c, \Lambda_{X|Y}) \end{aligned}$$

with the following formulas for the conditional mechanism

$$C = B^\top \tag{4.62}$$

$$c = \eta_X - B^\top \Lambda_{Y|X}^{-1} b \tag{4.63}$$

$$\Lambda_{X|Y} = \Lambda_X + B^\top \Lambda_{Y|X}^{-1} B \tag{4.64}$$

followed by these formulas for the marginal mechanisms

$$\Lambda_Y + C^\top \Lambda_{X|Y}^{-1} C = \Lambda_{Y|X} \tag{4.65}$$

$$\eta_Y - C^\top \Lambda_{X|Y}^{-1} c = b\,. \tag{4.66}$$

These formulas are going to be very useful to establish a relationship between the distance to optimum of the causal and anticausal models in Appendix 4.E.5.

4.E.3 Cholesky Parameters

We call Cholesky parametrization of the normal law the parameters $(\boldsymbol{L}, \zeta)$ such that

$$\begin{aligned} \Lambda &= \boldsymbol{L}\boldsymbol{L}^\top \qquad (\boldsymbol{L} \textit{ is lower triangular}) \\ \zeta &= \boldsymbol{L}^{-1}\eta = \boldsymbol{L}^\top \mu \end{aligned}$$

We use $\mathcal{N}_{\text{cho}}(\zeta, \boldsymbol{L})$ to denote the normal law with Cholesky parameters ζ and $\boldsymbol{L}$. The full causal model (4.61) becomes

$$\begin{aligned} X &\sim \mathcal{N}_{\text{cho}}(\zeta_X, \boldsymbol{L}_X) \\ Y|X &\sim \mathcal{N}_{\text{cho}}(MX + m, \boldsymbol{L}_{Y|X}) \end{aligned} \tag{4.67}$$

where the 5 parameters are defined from the natural model by the equations

$$\begin{aligned} \boldsymbol{L}_X L_X^\top &= \Lambda_X \\ \boldsymbol{L}_{Y|X}\boldsymbol{L}_{Y|X}^\top &= \Lambda_{Y|X} \\ \zeta_X &= \boldsymbol{L}_X^{-1}\eta_X \\ M &= \boldsymbol{L}_{Y|X}^{-1} B \\ m &= \boldsymbol{L}_{Y|X}^{-1} b \end{aligned} \tag{4.68}$$

There is no closed formula to express the Cholesky decomposition of a sum of matrix $A + B$ with the Cholesky decomposition of A and B. As a consequence, there is no simple formula to switch between causal and anticausal models with this parametrization.

Joint Cholesky

We derive a formula for the joint Cholesky for future reference. To get a closed form for the joint Cholesky parameters from the conditional parameters, we need to switch the positions of X and Y in the joint vector – e.g. we are using (Y, X) instead of (X, Y). Indeed the Cholesky decomposition is very dependent on the orders of the rows and columns. That's also why we cannot simply switch the column orders in the joint representation.

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathcal{N}_{\text{cho}} \left(\begin{pmatrix} m \\ \zeta_X \end{pmatrix}, \begin{pmatrix} \boldsymbol{L}_{Y|X} & 0 \\ -M^\top & \boldsymbol{L}_X \end{pmatrix} \right) .$$

This joint representation is simply taking the conditional parameters and putting them in an array. It has the advantage that the distance is equal to the conditional distance. It hints towards the idea that for multivariate normal variables, knowing the right Cholesky decomposition is equivalent to knowing the right causal graph.

4.E.4 Kullback-Leibler Divergence

We express the KL divergence in all three parametrizations because we use them in the code.

$$\begin{aligned} &2D_{\text{KL}}(\mathcal{N}_0||\mathcal{N}_1) \\ &= (\mu_1 - \mu_0)^\top \Sigma_1^{-1} (\mu_1 - \mu_0) \\ &\quad + \text{Tr}(\Sigma_1^{-1}\Sigma_0) - k - \log\left|\Sigma_1^{-1}\Sigma_0\right| \\ &= \eta_1^\top \Lambda_1^{-1} \eta_1 - 2\eta_1^\top \Lambda_0^{-1}\eta_0 + \eta_0^\top \Lambda_0^{-1}\Lambda_1\Lambda_0^{-1}\eta_0 \\ &\quad + \text{Tr}(\Lambda_1\Lambda_0^{-1}) - k - \log(\left|\Lambda_1\Lambda_0^{-1}\right| \\ &= \left\|V^\top \zeta_0 - \zeta_1\right\|^2 + \|V\|_F^2 - k - 2\log|V| \end{aligned}$$

where $V := \boldsymbol{L}_0^{-1}\boldsymbol{L}_1$ is a lower triangular matrix which plays a special role.

4.E.5 Distance after Intervention

In this section we evaluate the effect on interventions on the cause and effect for both models. When the intervention happens on the cause, we replace μ_X, Σ_X by $\tilde{\mu}_X, \tilde{\Sigma}_X$, or equivalently we replace the natural parameters of the marginal on X. The natural causal distance is simply

$$\delta_{\text{causal}} = \|\eta_X - \tilde{\eta}_X\|^2 + \|\Lambda_X - \tilde{\Lambda}_X\|_F^2 . \tag{4.69}$$

Unless indicated otherwise, we will consider the Frobenius distance between matrices. For the anticausal model, both marginal and conditional parameters need to change. Here similar to categorical case, we have

$$\delta_{\text{anticausal}} \geq \delta_{\text{causal}}.$$

However when the intervention happens on the effect, there is no clear formal relation between δ_{causal} and $\delta_{\text{anticausal}}$. More detail about the deriving the mathematical formula for δ_{causal} and $\delta_{\text{anticausal}}$ is presented in the following.

Intervention on Cause

When the intervention happens on the cause, the natural causal distance is

$$\delta_{\text{causal}} = \|\eta_X - \tilde{\eta}_X\|^2 + \|\Lambda_X - \tilde{\Lambda}_X\|_F^2 \,. \tag{4.70}$$

How does this transformation affect the anticausal parameters? Both the marginal and the conditional have to adapt. The anticausal conditional is elegantly expressed with the causal natural parameters in (4.64), so we will start with the conditional

$$C - \tilde{C} = B^\top - \tilde{B}^\top = 0 \tag{4.71}$$

$$c - \tilde{c} = \eta_X - \tilde{\eta}_X \tag{4.72}$$

$$\Lambda_{X|Y} - \tilde{\Lambda}_{X|Y} = \Lambda_X - \tilde{\Lambda}_X \,. \tag{4.73}$$

In words, the linear transformation is invariant, the bias moves like the mean of X, and the conditional precision moves like the precision of X. This means that we can directly lower bound the anticausal distance with the causal distance

$$\begin{aligned}
\delta_{\text{anticausal}} &= \left\|C - \tilde{C}\right\|^2 + \|c - \tilde{c}\|^2 + \left\|\Lambda_{X|Y} - \tilde{\Lambda}_{X|Y}\right\|^2 \\
&\quad + \left\|\Lambda_Y - \tilde{\Lambda}_Y\right\|^2 + \|\eta_Y - \tilde{\eta}_Y\|^2 \\
&= \delta_{\text{causal}} + \left\|\Lambda_Y - \tilde{\Lambda}_Y\right\|^2 + \|\eta_Y - \tilde{\eta}_Y\|^2 \\
&> \delta_{\text{causal}} \,.
\end{aligned} \tag{4.74}$$

We could get a stronger bound by bounding the anticausal marginal parameters, but any such bound would involve the value of the linearity B and make the result needlessly more complicated.

Intervention on Effect

We perform an intervention on Y such that the causal model become independent

$$\begin{aligned}
X &\sim \mathcal{N}(\mu_X, \Sigma_X) \\
Y &\sim \mathcal{N}(\tilde{\mu}_Y, \tilde{\Sigma}_Y) \,.
\end{aligned}$$

This independence means that the linear models have a slope 0

$$\tilde{A} = \tilde{B} = \tilde{C} = 0 \,. \tag{4.75}$$

The bias then has to account for the mean parameter

$$\tilde{a} = \tilde{\mu}_Y, \quad \tilde{b} = \tilde{\eta}_Y, \quad \tilde{c} = \eta_X \tag{4.76}$$

And the conditional precision have to match the marginal precision

$$\tilde{\Sigma}_{Y|X} = \tilde{\Sigma}_Y, \quad \tilde{\Lambda}_{Y|X} = \tilde{\Lambda}_Y, \quad \tilde{\Lambda}_{X|Y} = \Lambda_X \tag{4.77}$$

So the distances are written

$$\begin{aligned}
\delta_{\text{causal}} &= \|B\|_F^2 + \|b - \tilde{\eta}_Y\|^2 + \|\Lambda_{Y|X} - \tilde{\Lambda}_Y\|_F^2 \\
\delta_{\text{anticausal}} &= \|C\|_F^2 + \|c - \eta_X\|^2 + \|\Lambda_{X|Y} - \Lambda_X\|_F^2 \\
&\quad + \|\eta_Y - \tilde{\eta}_Y\|^2 + \|\Lambda_Y - \tilde{\Lambda}_Y\|_F^2 \\
&= \|B\|_F^2 + \|B^\top \Lambda_{Y|X}^{-1} b\|^2 + \|B^\top \Lambda_{Y|X}^{-1} B\|_F^2 \\
&\quad + \|\eta_Y - \tilde{\eta}_Y\|^2 + \|\Lambda_Y - \tilde{\Lambda}_Y\|_F^2
\end{aligned}$$

We did not find any meaningful simplification of these formulas.

4.F Normal Prior

Exactly like in the categorical setting, the distributions we sample are going to impact the speed of adaptation and the distances we measure. Let K be the dimension of X and Y, and $n_0 = 2K + 2$ an arbitrary number of prior observations. We define a *pseudo-conjugate prior*

$$\Lambda_X \sim \mathcal{W}(n_0, \frac{I_K}{K}) \tag{4.78}$$

$$\eta_X | \Lambda_X \sim \mathcal{N}(0, \frac{\Lambda_X}{n_0}) \tag{4.79}$$

$$\Lambda_{Y|X} \sim \mathcal{W}(n_0, 10\frac{I_K}{K}) \tag{4.80}$$

$$b | \Lambda_{Y|X} \sim \mathcal{N}(0, \frac{\Lambda_{Y|X}}{n_0}) \tag{4.81}$$

$$B = \Lambda_{Y|X} A \text{ where } A \sim \mathcal{N}(0, \frac{I_{K^2}}{\sqrt{K}}) \tag{4.82}$$

where $\mathcal{W}$ is the Wishart distribution with parameters: degrees of freedom and scale matrix. We picked these parameters such that η_Y, Λ_Y follows approximately the

same law as η_X, Λ_X. Two important factors to get a symmetric relationship between X and Y are 10 and $\sqrt{K}$. First, we sample a larger conditional precision, so that their relationship is quite deterministic. Second we sample the linear layer such that it preserves the scale of X, so that X and Y have approximately the same variance. We also use appropriate covariance matrices to sample other parameters such that the prior is somewhat conjugate and gives proper variance formulas.

We sample interventions from the same distributions as the cause marginal in (4.82). For an intervention on the cause

$$\begin{aligned} \tilde{\Lambda}_X &\sim \mathcal{W}(n_0, \frac{I_K}{K}) \\ \tilde{\eta}_X | \tilde{\Lambda}_X &\sim \mathcal{N}(0, \frac{\tilde{\Lambda}_X}{n_0}) \ . \end{aligned}$$

For an intervention on the effect

$$\begin{aligned} \tilde{B} = 0 \tilde{\Lambda}_{Y|X} = \tilde{\Lambda}_Y &\sim \mathcal{W}(n_0, \frac{I_K}{K}) \\ \tilde{b} = \tilde{\eta}_Y | \tilde{\Lambda}_Y &\sim \mathcal{N}(0, \frac{\tilde{\Lambda}_Y}{n_0}) \ . \end{aligned}$$

5 Convergence Rates for the MAP of an Exponential Family and Stochastic Mirror Descent – an Open Problem

Prologue to the Third Contribution

Article Details

Convergence Rates for the MAP of an Exponential Family and Stochastic Mirror Descent – an Open Problem. *Rémi Le Priol, Frederik Kunstner, Damien Scieur, and Simon Lacoste-Julien.* This paper is under review at AISTATS 2022 (Le Priol et al., 2021b).

History of this Paper

Rémi and Simon started looking for MAP convergence rates when he was involved in Le Priol et al. (2021a), as such rates may have been helpful to characterize the speed of adaptation of categorical of Gaussian models (see the 2nd contribution). About the same time, Frederik independently got interested in this problem while working on Kunstner et al. (2021), a beautiful paper proving that expectation-maximization (EM) in exponential families is an instance of mirror descent. They were then able to leverage recent results on relative smoothness (Lu et al., 2018) to get the first global convergence rate applicable to EM for gaussian mixture models. This result naturally raised the question : can we get similar results for stochastic EM via a convergence rate on stochastic mirror descent ? Meanwhile, Remi and Simon sent calls to the community, looking for answers to this seemingly simple question. That is when the trajectories of Rémi and Frederik collided, thanks to the mediation of Mark Schmidt. Damien finally hopped in on the project, bringing his optimization expertise. Together, they created this article, gathering all the takes they could find on this problem.

Contributions of the Authors

Rémi Le Priol came up with the equivalence between MAP and SMD, wrote most of the paper, and made the figures. Frederik Kunstner contributed to the literature review, and in particular the comparison between analysis of SMD. Damien Scieur contributed to the general writing and the results on self-concordance. Damien Scieur and Simon Lacoste-Julien provided supervision.

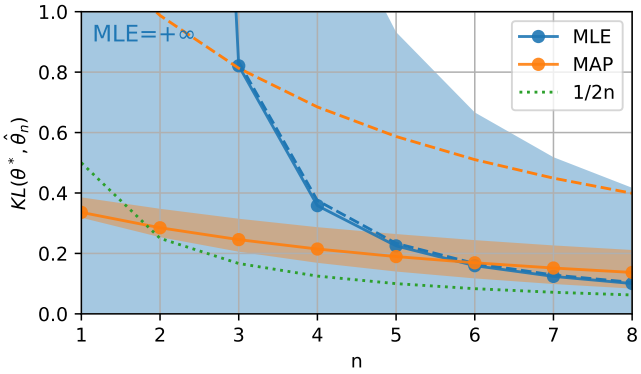


Figure 5.1 – KL divergence (5.4) for Gaussian variance (§5.4.1) MLE (blue) and MAP (orange) against number of samples n. Solid curve are average over 10^5 trials. Dashed curves are upper bounds (5.18) (blue) and (5.19) (orange, not tight by a factor 2). Shaded areas are 90% confidence interval. The MLE expected KL is infinite for $n = 1$ and $n = 2$, but for $n \geq 3$ it quickly joins the upper bound (5.18) and the $1/2n$ asymptote (5.23). MAP's expected KL is always finite, and it has lower variance than MLE, but it is slower to join the asymptote. We wish to find upper bounds similar to (5.19) characterizing the relative importance of the prior and the few sample behavior of MAP for a variety of exponential families.

Abstract

We consider the problem of upper bounding the expected log-likelihood sub-optimality of the maximum likelihood estimate (MLE), or a conjugate maximum a posteriori (MAP) for an exponential family, in a non-asymptotic way. Surprisingly, we found no general solution to this problem in the literature. In particular, current theories do not hold for a Gaussian or in the interesting few samples regime. After exhibiting various facets of the problem, we show we can interpret the MAP as running stochastic mirror descent (SMD) on the log-likelihood. However, modern convergence results do not apply for standard examples of the exponential family, highlighting holes in the convergence literature. We believe solving this very fundamental problem may bring progress to both the statistics and optimization communities.

5.1 Introduction

Models Exponential families are among the most widely used simple parametric models of data, yet, we will highlight some open problems about them in this paper. Many standard random variables are exponential families: Gaussians, categorical, gamma, or Dirichlet, for example. They are flexible enough to model a variety

of data sources X and easy to describe with some sufficient statistics $T(X) \in \mathbb{R}^d$. They are particularly appreciated for their convex log-likelihood

$$f(\theta) := \mathbb{E}[-\log p_\theta(X)] = A(\theta) - \langle \mathbb{E}[T(X)], \theta \rangle, \tag{5.1}$$

where A is the convex log-partition function and $\theta \in \Theta$ is the *natural parameter*. This convexity lays the foundation for generalized linear models (McCullagh and Nelder, 1989) or variants of principal component analysis (Collins et al., 2001), among other applications.

Estimators In this paper, we consider the problem of estimating θ from a dataset $\boldsymbol{D} = (X_1, \dots, X_n)$ of iid observations from p_θ in an exponential family. In this case, not only is f convex, but it yields a simple condition for the maximum-likelihood estimate (MLE)

$$\hat{\mu}_n^{\text{MLE}} = \nabla A(\hat{\theta}_n^{\text{MLE}}) = \frac{\sum_{i=1}^n T(x_i)}{n} \ . \tag{5.2}$$

This rule is also known as moment matching. Given a specific conjugate prior, a similar formula (5.11) holds for the maximum a posteriori (MAP). In this paper, we will focus on analyzing MLE and MAP estimators.[1]

Statistical decision theory To assess the quality of an estimator $\hat{\theta}$ (and compare them), we need to define some notion of closeness to the correct parameter θ^*. We distinguish here two ways: *distance in parameter space* and *"distance" between distributions*. **1)** Distance in parameter space $d(\theta, \theta^*)$. This is the focus of *parameter estimation*, yielding results such as the asymptotic efficiency of the MLE via the Cramer-Rao lower-bound (Aitken and Silverstone, 1942) and a wealth of asymptotic results (Van der Vaart, 1998). In particular for sum of independent variables such as (5.2), large deviations theory (Varadhan, 1984) characterizes concentration phenomena. **2)** Distance between distributions, as studied in *density estimation*. For this purpose, the Kullback-Leibler (KL) divergence $D_{\text{KL}}(p_{\theta^*}||p_\theta)$ arises naturally from information theory, but its lack of robustness to misspecification[2] has led statisticians to study better-behaved distances, such as the L^2 norm (Tsybakov, 2009, §1.2), the L^1 norm (Devroye and Lugosi, 2001), or more recently χ^2 distance (Kamath et al., 2015) or Hellinger distance (Baraud et al., 2017). With exponential families, the KL divergence is also a Bregman divergence between parameters (see §5.3), thus drawing a connection between these two lines of research, and raising

[1] A related analysis is present in the online-learning literature, but for different online estimators, which are less efficient than offline methods (Azoury and Warmuth, 2001; Dasgupta and Hsu, 2007).

[2] For p and q continuous densities, $D_{\text{KL}}(p||q) = +\infty$ if $\exists x, q(x) = 0 \,\&\, p(x) > 0$.

the fundamental problem:

$$\boxed{\begin{array}{c}\textit{Find an upper bound on the expected value of} \\ D_{\mathrm{KL}}(p_{\theta^*}||p_{\hat{\theta}_n^{\mathrm{MLE/MAP}}})\ .\end{array}} \qquad (\star)$$

There are already general asymptotic results (§5.5.1 and Fig. 5.1), and a finite n result when A is quadratic (e.g., X is Gaussian with known variance) or close to quadratic (§5.5.2). However, a general solution for finite n remains elusive. In this paper, we review partial solutions and give ideas on how to solve the problem.

Optimization Stochastic optimization offers an interesting perspective on ($\star$). Consider the problem

$$\min_{\theta\in\Theta} f(\theta)\,, \qquad (5.3)$$

solved by $\theta^* \in \Theta$. Setting f as the log-likelihood (5.1), the suboptimality is equal to the KL:

$$f(\theta) - f(\theta^*) = D_{\mathrm{KL}}(p_{\theta^*}||p_\theta). \qquad (5.4)$$

Both MLE and MAP can be seen as stochastic algorithms solving (5.3). In particular, with exponential families, MAP is equivalent to stochastic mirror descent (SMD) (Nemirovski et al., 2009). Inspired by recent work (Le Priol et al., 2021a; Kunstner et al., 2021), we consider using existing convergence rates for SMD to get the upper bound we seek. Unfortunately, none of the current analyses apply, highlighting open problems for the analysis of SMD.

Expected Outcomes A solution to ($\star$) can clarify the importance of the prior in MAP, in particular in the few sample regime. Also, it could enable stochastic optimization to tackle a broad class of barrier objectives.[3] A good example is the generalized linear model based on Gaussians with unknown mean and variance, for which there is currently no theory (Bach and Moulines, 2013). It could also help assess the impact of alternative forms of regularization (prior) for these models.

Contributions After formalizing the problem ($\star$) (§5.3), along with its asymptotic properties (§5.5.1), we make the following contributions.

- We provide an upper bound on the KL in the particular case of a Gaussian with known mean but unknown variance $\mathcal{N}(0, \sigma^2)$ (§5.4.1), illustrating that tight rates are possible even though the current theory does not cover them.

[3] we call *barrier* an objective f that is infinite on the boundaries of its domain (assuming they exist).

- We highlight sufficient conditions to characterize when a (local) quadratic approximation of the KL is valid, offering a partial answer to ($\star$) (§5.5.2-5.5.3).
- By linking MAP and SMD, we show that modern analysis of SMD is yet to prove convergence on barrier objectives such as $-\log$ (§5.6).

Notation X and $T = T(X)$ are random variables, x is a sample, n is the number of samples and $d = \dim(T)$. $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product in $\mathbb{R}^d$.

5.2 Technical Background

This section reviews the formalism of exponential families, their duality, a conjugate prior, and the corresponding MAP. We point the reader towards Wainwright and Jordan (2008, Chapter 3) for a more detailed introduction.

The density of an exponential family for a sample x is

$$p_\theta(x) = p(x|\theta) = \exp(\langle \theta, T(x) \rangle - A(\theta)) \ , \tag{5.5}$$

where θ is called natural (or primal) parameter. It is fully specified by 1) $T : \mathcal{X} \to \mathbb{R}^d$, the sufficient statistic, and 2) a base measure ν on $\mathcal{X}$. Since the exponential is positive, p has the same support as ν. The log-partition function A acts as a normalization term, since

$$A(\theta) = \log \int e^{\langle \theta, T(x) \rangle} \nu(dx) \ . \tag{5.6}$$

This simple model encompasses both categorical distributions : $\mathcal{X} = \{1, \dots, k\}$, ν uniform and $T(X)$ the one-hot encoding and multivariate normal distributions $\mathcal{X} = \mathbb{R}^d, \nu$ Lebesgue and $T(X) = (X, XX^\top)$.

For convenience, we focus on steep, regular exponential families with minimal statistic T (Barndorff-Nielsen, 1978). Then A is a strictly convex function of Legendre type, and the set $\Theta = \{\theta \,|\, A(\theta) < \infty\}$ is open and convex. When explicit, we write the random variable $T = T(X)$.

Duality. The log-partition function A verifies the two following identities:

$$\nabla A(\theta) = \mathbb{E}_{p_\theta}\left[T(X)\right] =: \mu, \tag{5.7}$$
$$\nabla^2 A(\theta) = \mathrm{Cov}_{p_\theta}[T(X)] > 0, \tag{5.8}$$

where μ is called the mean (or dual) parameter, which lives in the open convex set $\mathcal{M}$ equal to the relative interior of the convex hull of $T(\mathcal{X})$. Given that A is strictly convex, its Hessian is positive definite, and its gradient ∇A is a *bijection* between natural parameters θ and mean parameters μ. We will write μ or θ interchangeably

depending on the context, being aware that both are linked and represent the same distribution.

We now introduce the convex conjugate (the Fenchel-Legendre transform) of the log-partition function

$$A^*(\mu) = \langle \mu, \theta \rangle - A(\theta) = \max_{\theta' \in \Theta} \langle \mu, \theta' \rangle - A(\theta') \ ,$$

which is the common notion of *entropy* in information theory. By Fenchel duality, its gradient is the inverse of the gradient of A, $\nabla A^* = \nabla A^{-1}$, giving

$$\nabla A^* \circ \nabla A(\theta) = \theta, \quad \nabla A \circ \nabla A^*(\mu) = \mu.$$

The Bregman Divergence induced by A measures the discrepancy between two parameters θ and θ_0,

$$\mathcal{B}_A(\theta; \theta_0) = A(\theta) - A(\theta_0) - \langle \nabla A(\theta_0), \theta - \theta_0 \rangle, \tag{5.9}$$

with $\nabla A(\theta_0) = \mathbb{E}_{\theta_0}[T(X)] =: \mu_0$ the mean parameter associated to θ_0. In general, Bregman divergences are not symmetric, i.e., $\mathcal{B}_A(\theta; \theta_0) \neq \mathcal{B}_A(\theta_0; \theta)$.

A Conjugate Prior for $p(X|\theta)$ is

$$\begin{aligned} p(\theta) &\propto \exp(-n_0 \mathcal{B}_A(\theta; \theta_0)) \\ &\propto \exp(n_0 \langle \mu_0, \theta \rangle - n_0 A(\theta)), \end{aligned} \tag{5.10}$$

where n_0 and θ_0 are (hyper)parameters of the prior (Agarwal and Daumé, 2010). This is an exponential family with sufficient statistics $(\theta, A(\theta))$ and natural parameter $(n_0\mu_0, -n_0)$. Intuitively, n_0 is the number of fictive data points observed from a distribution with natural parameter θ_0.

Maximum A Posteriori (MAP). Given a dataset $\boldsymbol{D}_n = (X_1, \ldots, X_n)$, we wish to estimate the maximum of the posterior distribution $p(\theta \mid \boldsymbol{D}_n) \propto p(\boldsymbol{D}_n|\theta)p(\theta)$. Plugging in (5.5), (5.9) and (5.10) yields

$$p(\theta \mid \boldsymbol{D}_n) \propto \exp(-(n_0 + n)\mathcal{B}_A(\theta; \hat{\theta}_n^{\text{MAP}}))$$

which reaches its maximum at $\hat{\theta}_n^{\text{MAP}}$ such that

$$\nabla A(\hat{\theta}_n^{\text{MAP}}) = \hat{\mu}_n^{\text{MAP}} = \frac{n_0\mu_0 + \sum_{i=1}^n T_i}{n_0 + n} \ , \tag{5.11}$$

where $T_i = T(X_i)$. When $n_0 = 0$ (no samples from the prior), we recover the MLE (5.2). We write $\hat{\theta}_n$ for the MAP and view the MLE as a particular case.

5.3 Problems Formulation

We are now ready to formalize the main problem of this paper. Assume we observe a dataset $\boldsymbol{D}_n$ drawn i.i.d. from $p(\cdot \,|\, \theta^*)$, an exponential family distribution with parameters θ^*. We wish to quantify how well the MLE or a MAP approximates the true distribution.

A natural way to quantify this is the Kullback-Leibler divergence (KL) $D_{\mathrm{KL}}(p_{\theta^*}||p_\theta)$. In the well-specified setting, it corresponds to the log-likelihood sub-optimality (5.4). With exponential families, the KL is also a Bregman divergence:

$$D_{\mathrm{KL}}(p_{\theta^*}||p_\theta) = \mathcal{B}_A(\theta; \theta^*) = \mathcal{B}_{A^*}(\mu^*; \mu) \; . \tag{5.12}$$

The second equality is a general property of Bregman divergences and convex conjugates. How does this quantity behave when $\hat{\theta}$ is the MLE or MAP? Or in the words of statistical decision theory, what is the *frequentist risk* of these estimators when the loss is the KL divergence? This is our first problem.

Open Problem 1 (Upper-bounding MAP and MLE)**.** *Upper bound the following quantities:*

$$\textit{MLE:} \quad \mathbb{E}_{\boldsymbol{D}_n}\left[\mathcal{B}_{A^*}\left(\mathbb{E}_{\theta^*}[T]; \frac{1}{n}\textstyle\sum_i T_i\right)\right], \tag{5.13}$$

$$\textit{MAP:} \quad \mathbb{E}_{\boldsymbol{D}_n}\left[\mathcal{B}_{A^*}\left(\mathbb{E}_{\theta^*}[T]; \frac{n_0\mu_0 + \sum_i T_i}{n_0 + n}\right)\right], \tag{5.14}$$

where the expectation is on the data $\boldsymbol{D}_n = (T_1, \dots, T_n)$.

More explicitly, we want an upper bound that does not involve this expectation over the dataset. Surprisingly, we found no general solution to this seemingly simple problem, whether in the literature or by our means. In §5.4, we provide results for special cases such as $\mathcal{N}(0, \sigma^2)$, while in §,5.5 we provide realistic conditions to obtain valid bounds after seeing a large number of samples. However, we have yet to find a solution encompassing both a broad range of exponential families and applicable to small sample sizes $n \lesssim d$.

A Difficulty with the MLE. While (5.14) is always finite, (5.13) may be infinite, for instance when estimating the covariance of a Gaussian when $n \leq d + 1$. Even worse, there is a non-zero probability never to sample one of the categories with categorical variables. In those cases the MLE gives zero weight to this category and $D_{\mathrm{KL}}(p_{\theta^*} \,\|\, p_{\mathrm{MLE}}) = +\infty$. Therefore, the expected KL (5.13) is infinite for any number of samples. Instead of taking the expectation, one might want to bound the risk in high probability without resorting to Markov inequality, as achieved by (Ostrovskii and Bach, 2021), but this is a difficult endeavor. These examples make a case for regularized estimators such as MAP, for which we may find upper bounds.

Optimization. With exponential families, MAP can be linked to *stochastic mirror descent (SMD)*, see App. 5.F. More precisely, let us re-write (5.11) as

$$\mu_n = \mu_{n-1} - \gamma_n(\mu_{n-1} - T_n) \tag{5.15}$$

where $\gamma_n := \frac{1}{n_0+n}$. Now define stochastic functions $f_X(\theta) = -\log p(X \mid \theta)$ such that $\mathbb{E}[f_X] = f$. If we further introduce stochastic gradients $g_n(\theta) := \nabla A(\theta) - T_n = \nabla f_{X_n}(\theta)$, then (5.15) becomes

$$\nabla A^*(\hat{\theta}_n) = \nabla A^*(\hat{\theta}_{n-1}) - \gamma_n g_n(\hat{\theta}_{n-1}), \tag{5.16}$$

which is the update formula for SMD on f with mirror map ∇A and step-size schedule γ_n, initialized at θ_0. In this view, MLE forgets its (arbitrary) initialization after the first step with step size 1. The observation MAP $\in$ SMD brings us to our second problem.

Open Problem 2 (Convergence rate for SMD)**.** *Find a convergence rate for stochastic mirror descent that applies to conjugate MAP of exponential families such as Gaussians $\mathcal{N}(\mu, \sigma^2)$.*

To address these problems, we start by investigating simple examples to provide solutions to Problem 1, getting insights into what is achievable.

5.4 Illustrating Examples

5.4.1 Gaussian with Unknown Variance

A non-trivial yet straightforward example is the centered Gaussian distribution with unknown variance $\mathcal{N}(0, \sigma^2)$. Its log-likelihood reads $\log p(x) = -\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$. Defining $T(X) = X^2$ as the sufficient statistic, we get natural parameter $\theta = -\frac{1}{2\sigma^2} < 0$, and mean parameter $\mu = \mathbb{E}[T(X)] = \sigma^2 > 0$. Mean and natural parameters are roughly inverse of each other, i.e., $\theta = -\frac{1}{2\mu}$. Now we match the log-likelihood with the exponential family template to get the log-partition function, and take the conjugate to find the entropy

$$A(\theta) = -\frac{1}{2}\log(-\theta) \quad \text{and} \quad A^*(\mu) = -\frac{1}{2}\log(\mu)\ ,$$

up to constants. Both A and the entropy are roughly negative logarithm $z \mapsto -\log(z)$. It means the conjugate prior is the exponential family with sufficient statistic $(\theta, \log(-\theta))$, e.g., a negative gamma distribution. It also means $\mathcal{B}_A$ and $\mathcal{B}_{A^*}$ have the same shape

$$\mathcal{B}_{A^*}(\mu^*; \mu_n) = \frac{1}{2}\left(\frac{\mu^*}{\mu_n} - 1 - \log\frac{\mu^*}{\mu_n}\right)\ . \tag{5.17}$$

In Theorems 5.4.1 and 5.4.2, we report upper bounds on the expected value of this divergence for the MLE and the MAP. All proofs for this section are in App. 5.A.

Theorem 5.4.1 (MLE Bound)**.** *The MLE of* $\mathcal{N}(0, \mu^*)$ *is* $\hat{\mu}_n^{MLE} = \frac{1}{n}\sum_i X_i^2$*. Its expected suboptimality is infinite when* $n \leq 2$*, and otherwise upper-bounded as*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n^{MLE})\right] \leq \frac{1}{2n} + \frac{2}{n(n-2)} \,. \tag{5.18}$$

This upper bound matches the asymptotic result (5.23) that we derive in §5.5.1. We illustrate its numerical behavior in Figure 5.1. With the same technique, we obtain a similar bound for the multivariate generalization: the expected value is infinite whenever $n \leq d+1$ where d is the dimension, and is otherwise bounded by $O(\frac{d^2}{n} + \frac{d^3}{n(n-d-1)})$.

Theorem 5.4.2 (MAP Bound)**.** *The expected suboptimality of the MAP of* $\mathcal{N}(0, \mu^*)$ *with prior hyper-parameters* (n_0, μ_0) *is*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n^{\text{MAP}})\right] \leq \begin{cases} \frac{1}{2(n_0+1)} + b_1 & \text{if } n = 1, \\ \frac{1}{n_0 \frac{\mu_0}{\mu^*} + n - 2} + b_n & \text{if } n \geq 2 \end{cases} \tag{5.19}$$

$$\text{where } b_n = \frac{(1 + \frac{1}{n_0} - \frac{\mu_0}{\mu^*})^2}{2(\frac{\mu_0}{\mu^*} + \frac{\max(0, n-2)}{n_0})(1 + \frac{n}{n_0})} \,.$$

Anticipating on §5.5.4, this inequality highlights an explicit $O(\frac{v}{n} + \frac{b}{n^2})$ variance-bias decomposition. This inequality is derived with the symmetrized Bregman $\mathcal{B}(a, b) + \mathcal{B}(b, a)$ for which calculus is more tractable. This explains why the variance term is twice larger than the asymptote (5.23). Regarding the bias, it vanishes when $\frac{\mu_0}{\mu^*} = 1 + \frac{1}{n_0}$, which happens when the prior is slightly larger than the ground truth. This correlates well with our numerical observations (cf App. 5.A).

Note that if $X \sim \mathcal{N}(0, \sigma^2)$, then $X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2\sigma^2})$ in the shape-rate parametrization of Gamma distributions. In fact the bounds above can be generalized to any distribution $\Gamma(\alpha, \beta)$ with known shape α. This generalization encompasses exponential distribution when $\alpha = 1$, as another important special case. We postpone these rates to Section 5.A for the sake of clarity.

5.4.2 Full Gaussian (Non-Trivial)

Now that we have solved the case of $\mathcal{N}(0, \sigma^2)$, consider the full Gaussian $\mathcal{N}(m, \sigma^2)$, which offers a highly non-trivial example for Problem 1. Their log-likelihood reads $p(x) = -\frac{(x-m)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$. With sufficient statistic $T(x) = (x, x^2)$, the mean parameters are $\mu = \mathbb{E}[T(X)] = (m, m^2 + \sigma^2)$ belonging to the open set $\mathcal{M} = \{(u, v) \,|\, u^2 < v\}$, and the natural parameters are $\theta = (\frac{m}{\sigma^2}, \frac{-1}{2\sigma^2}) \in \Theta =$

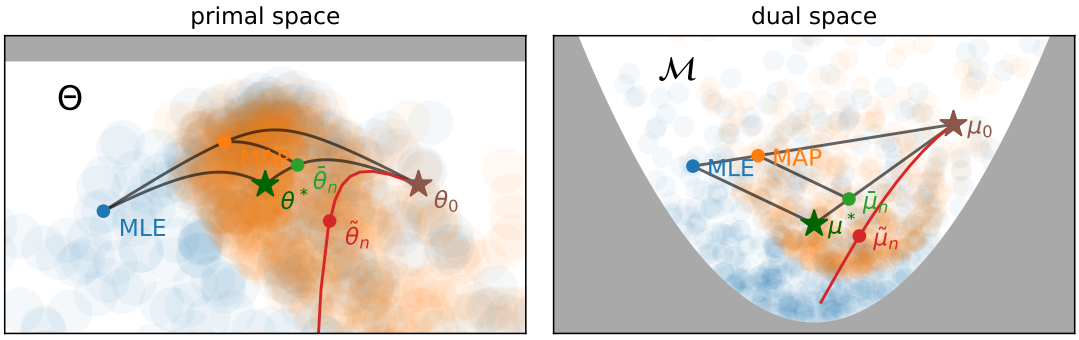


Figure 5.2 – Primal and dual representations of a Gaussian $\mathcal{N}(m, \sigma^2)$ MLE (blue) and MAP (orange) (§5.4.2 with $n = 3$). In dual space, MAP is a scaled version of the MLE (5.11) with expectation $\mathbb{E}[\hat{\mu}_n^{\text{MAP}}] =: \bar{\mu}_n$ (light green), and MLE is unbiased $\mathbb{E}[\hat{\mu}_n^{\text{MLE}}] = \mu^*$, as illustrated by the parallels in the grey triangle. In primal space, MAP has expectation $\tilde{\theta}_n$ (red), which intervenes in the bias-variance decomposition (5.30) from §5.5.4. The hyperparameter of the prior θ_0 controls the brown point's location while varying n_0 spans the long edges of the triangle and the red curve. Large blurry circles in the background are other instances of MAP and MLE revealing their distribution.

$\mathbb{R} \times \mathbb{R}_-$. Examples of MAP and MLE are represented in Fig. 5.2 within $\mathcal{M}$ and Θ delimited in grey. Given these parameters, log-partition and entropy are, up to constants,

$$A(\theta) = \tfrac{\theta_1^2}{-4\theta_2} - \tfrac{1}{2}\log(-\theta_2) \tag{5.20}$$

$$A^*(\mu) = -\tfrac{1}{2}\log(\mu_2 - \mu_1^2) \tag{5.21}$$

These functions are neither smooth, nor strongly convex, but they are self-concordant, since A^* is the logarithmic barrier of a quadratic domain (Nesterov, 2004c, p.177, example 4.1.1.4), and self-concordance is preserved by convex-conjugacy (Nesterov and Nemirovskii, 1994) – see more details in App. 5.B. We now discuss the general problem and some ways to solve it via direct expansions of the Bregman divergence.

5.5 Partial Solutions

5.5.1 Asymptotic Rate

As a reference point for any finite convergence rate, it is interesting to briefly review the classical asymptotic behavior of these quantities as $n \to +\infty$. Proofs are in App. 5.C, and Ostrovskii and Bach (2021, §1.1) offers a more comprehensive review.

Statistics typically give results on θ, but the MAP (5.11) is more simply expressed with μ, so let us focus on $\mathcal{B}_{A^*}$. Bregman divergences are locally quadratic, as seen via a second order Taylor expansion

$$\mathcal{B}_{A^*}(\mu^*;\mu) = \tfrac{1}{2}\|\mu^* - \mu\|_{\boldsymbol{F}}^2 + O(\|\mu - \mu^*\|^3), \tag{5.22}$$

where the Mahalanobis norm $\|x\|_{\boldsymbol{F}}^2 = x^\top \boldsymbol{F} x$ is induced by $\boldsymbol{F} := \nabla^2 A^*(\mu^*)$, the Hessian of the entropy at the optimum. It happens that $\boldsymbol{F}$ is also the inverse *Fisher information matrix* at θ^*, since

$$\boldsymbol{F} := \nabla^2 A^*(\mu^*) = \nabla^2 A(\theta^*)^{-1} = \mathrm{Cov}_{\theta^*}[T(X)]^{-1} \ .$$

Plugging the MLE (5.2) or MAP (5.11) into (5.22), we get

$$\mathbb{E}\,\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)]; \hat{\mu}_n^{\mathrm{MLE/MAP}}\right) = \frac{d}{2n} + O(n^{-\frac{3}{2}}) \ . \tag{5.23}$$

Both MLE and MAP have the same asymptote, as the contribution of the prior $n_0\mu_0$ gets negligible for large n. This asymptote is independent of the optimum μ^* or $\boldsymbol{F}$ for well-specified models. Actually, $\frac{d}{2n}$ is also the minimax of the KL over all estimators, at least for categorical data (Braess and Sauer, 2004; Kamath et al., 2015). Next, we focus on the quadratic example, whose finite sample rate matches (5.23).

5.5.2 Quadratic Case

As another classical reference point, we consider the case $A(\theta) = \frac{1}{2}\|\theta\|_2^2$. For instance, this is the log-partition of a Gaussian with known variance I,

$$\mathcal{X} = \mathbb{R}^d, \quad \nu(dx) = \exp(\tfrac{-\|x\|^2}{2})dx, \quad T(x) = x.$$

In this case, $A^*(\mu) = \frac{1}{2}\|\mu\|_2^2$ as well, and both Bregman divergences are squared ℓ^2 distances since

$$\mathcal{B}_{A^*}(\mu^*;\mu) = \frac{1}{2}\|\mu^* - \mu\|_2^2 \ . \tag{5.24}$$

Thanks to the independence of samples, we can break down the MLE into individual point's contributions:

$$\mathbb{E}\left[\frac{1}{2}\left\|\mu^* - \frac{1}{n}\textstyle\sum_i T_i\right\|_2^2\right] = \frac{\mathrm{Var}(T)}{2n} = \frac{d}{2n}. \tag{5.25}$$

Adding a reference mean μ_0 to get the MAP yields

$$\mathbb{E}\left[\frac{1}{2}\left\|\mu^* - \hat{\mu}_n^{\mathrm{MAP}}\right\|_2^2\right] = \frac{n\,\mathrm{Var}(T) + n_0^2\|\mu^* - \mu_0\|^2}{2(n+n_0)^2} \ . \tag{5.26}$$

We see here a variance term defining the $\frac{d}{2n}$ asymptote and a bias term in $O(n^{-2})$. However, this result does not generalize well to other families unless we make restrictive assumptions on A^*.

If A^* is L-Lipschitz (e.g. A is defined within the ℓ^2-ball of radius L), then

$$\mathcal{B}_{A^*}(\mu^*;\mu) \leq L\|\mu^* - \mu\| + \|\theta\|\|\mu^* - \mu\| \tag{5.27}$$
$$\leq 2L\|\mu^* - \mu\| \,, \tag{5.28}$$

so $\mathcal{B}_{A^*}$ is Lipschitz, and (5.26) yields a $O(\frac{1}{\sqrt{n}})$ rate, but no common exponential families verify the assumption.

If A^* is L-smooth[4] (e.g. A is $\frac{1}{L}$-strongly convex (Kakade et al., 2009)), then

$$\mathcal{B}_{A^*}(\mu^*;\mu) \leq \frac{L}{2}\|\mu^* - \mu\|^2 \,, \tag{5.29}$$

so $\mathcal{B}_{A^*}$ is upper bounded by a quadratic, and we get (5.26) as an upper bound. It is also possible to get (more complex) upper bounds under restricted notions of strong-convexity (Negahban et al., 2012). Besides the Gaussian with known variance, the problem is that no standard exponential family has a *globally* strongly convex log-partition function. The next section focuses on *local* quadratic behavior, which is more realistic.

5.5.3 Locally Quadratic Case

From the Taylor expansion (5.22), we know that all Bregman divergences are locally quadratic. Under some assumptions, such as self-concordance[5] of A^* (Nesterov, 2004c, Ch. 4.1), we can quantify when this quadratic behavior kicks in. Proofs for this subsection are in App. 5.D.

Proposition 5.5.1. *Let $A^* : \mathcal{M} \to \mathbb{R}$ be a self-concordant convex function, $\mu, \mu^* \in \mathcal{M}$ and $\boldsymbol{F} = \nabla A^*(\mu^*)$. Then*[6]

$$\|\mu^* - \mu\|_{\boldsymbol{F}} < 0.21 \implies \mathcal{B}_{A^*}(\mu^*;\mu) \leq \|\mu^* - \mu\|_{\boldsymbol{F}}^2.$$

To gain insights into how many samples are needed, we can estimate when $\mathbb{E}[\|\mu^* - \mu\|_{\boldsymbol{F}}] < 0.21$. For the MLE, the proof of (5.23) from (5.22) yields $\mathbb{E}[\|\mu^* - \hat{\mu}_n\|_{\boldsymbol{F}}^2] = \frac{d}{n}$ in general, so a sufficient condition is $n \geq 25d$. For MAP, transforming (5.26), we get the sufficient condition $n \geq 25d + 5\|\mu^* - \mu_0\| - n_0$. This means that on average, we need 25 times more samples than the dimension to reach the quadratic regime and ensure an upper-bound like (5.26).

[4] A^* is L-smooth iff ∇A^* is L-Lipschitz.

[5] In 1d, f is self-concordant iff $\forall x, |f'''(x)| \leq 2|f''(x)|^{\frac{3}{2}}$.

[6] 0.21 is a value of x such that $x^2 \geq -\frac{x}{1-x} - \log(1 - \frac{x}{1-x})$.

The entropy A^* is self-concordant for several common families such as Gaussians (§5.4.2), and all families with $A \approx -\log$ such as exponential distributions, Laplace with known mean, Pareto with known minimum value, or Weibull with known shape k. The entropy is also self-concordant when T lives in a compact (Bubeck and Eldan, 2015) – e.g., categorical and Dirichlet distributions. Precisely, categorical variables illustrate that Proposition 5.5.1 does not imply directly a bound on the *expected* Bregman. The expected KL of the categorical MLE is always infinite, as previously mentioned in §5.3. However, Proposition 5.5.1 may be used to prove high-probability, many samples, convergence rates for one dimensional (and possibly multivariate) normal distributions.

This is the spirit of Ostrovskii and Bach (2021) which characterizes the number of samples needed to be upper bounded by a quadratic *with high-probability*, for any parametric models with a self-concordant log-likelihood f. Anastasiou and Reinert (2017) obtains a similar flavor of result under other assumptions on the third derivative of f. More closely, in the world of exponential families, Kakade et al. (2010) prove a result similar to Proposition 5.5.1 from a local bound on all higher-order moments of A in θ^*. However, these results are expressed with quadratics in θ, not μ, and they do not directly translate to convergence rates for the MAP, but they might with some more work.

More generally, the present proposition and these related works answer ($\star$) only *partially*, as they all give *large sample* results, that hold when $n \geq N$ for some constant N. A full solution to ($\star$) would apply to small n. Informed by the properties that we have seen so far, we next investigate a general decomposition of the Bregman that could guide us towards a solution.

5.5.4 Bias-Variance Decomposition

In both the quadratic (5.26) and the Gaussian variance examples (5.19), the upper bound takes the form $O(\frac{1}{n}) + O(\frac{\text{bias}}{n^2})$, giving us a flavor of what we would like as a general result for exponential families: a finite sample convergence rate, with variance and bias terms that reflect the important constants of the problem. Such a decomposition exists for any Bregman divergence (Pfau, 2013, Theorem 0.1).

Theorem 5.5.2 (Bregman Bias-Variance Decomposition)**.** *Let $\tilde{\theta}_n := \mathbb{E}[\hat{\theta}_n]$ be the expectation of the MAP in primal space, and $\tilde{\mu}_n = \nabla A(\tilde{\theta}_n)$ be its dual representation. The expected Bregman decomposes into*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n)\right] = \mathcal{B}_{A^*}(\mu^*; \tilde{\mu}_n) + \mathbb{E}\left[\mathcal{B}_{A^*}(\tilde{\mu}_n; \hat{\mu}_n)\right] \tag{5.30}$$

We plot this decomposition for $\mathcal{N}(\mu, \sigma^2)$ in Fig. 5.3 , and we illustrate the primal mean $\tilde{\theta}_n$ in Fig. 5.2.

Remark: In this decomposition, the primal expectation $\mathbb{E}[\hat{\theta}_n]$ is the reference point. An estimator will be unbiased if $\tilde{\theta}_n = \theta^*$. This is not true for the MLE, which is unbiased w.r.t. the dual parameter $\mathbb{E}[\hat{\mu}_n] = \mu^*$.

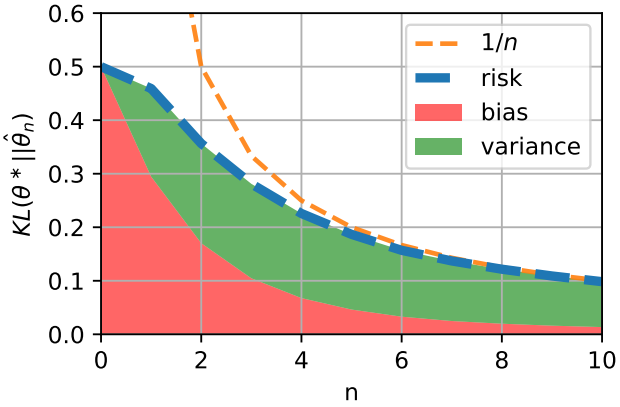


Figure 5.3 – Bias-Variance Decomposition for a Gaussian $\mathcal{N}(m, \sigma^2)$ with $\mu^* = (0, 1), \mu_0 = (1, 2)$ and $n_0 = 1$. The asymptote is $\frac{1}{n}$.

We show in App. 5.E that the bias decreases like $\mathcal{B}_{A^*}(\mu^*; \tilde{\mu}_n) \leq \frac{2}{n(n-2)}$ for Gaussian variance MLE, and $\mathcal{B}_{A^*}(\mu^*; \tilde{\mu}_n) \leq \frac{\|\mu^* - \mu_0\|^2}{(1+\frac{n}{n_0})^2}$ for a quadratic MAP. These observations hint towards a general $O(1/n^2)$ upper bound for the bias, while the variance may be less dependent on the initialization θ_0.

In this section, we considered direct expansions of (5.14). None of them could fully solve ($\star$). Next, we investigate whether an optimization approach could solve it.

5.6 An Optimization Problem

As we saw in §5.3, MAP can be interpreted as stochastic mirror descent (SMD). This means that **1)** we may obtain a convergence rate for MAP from an optimization analysis, and **2)** any insights gained from MAP may inform other designs and analyses of SMD. In particular, we know that MAP converges asymptotically as $O(n^{-1})$, so we hope to find a convergence rate for SMD that could capture this behavior. We first review the assumptions of relative smoothness, helpful to deal with non-smooth functions, before investigating recent analyses of SMD with the MAP.

5.6.1 Relative Smoothness

Mirror descent (MD) (Nemirosky and Yudin, 1983; Beck and Teboulle, 2003), also known as Bregman (proximal) gradient, relative gradient descent or NoLips, and SMD (Nemirovski et al., 2009; Ghadimi and Lan, 2012) are typically encountered in

Table 5.1 – Summary of results for SMD under relative smoothness and relative strong convexity assumptions. Each row correspond to one analysis, and each columns answers one question. $(-\log)$ does the bound hold for the Gaussian variance example (§5.4.1)? $(\gamma_n \sim \frac{1}{n})$ does it converge with a $O(\frac{1}{n})$ step-size? (f) is the bound in function value, or in reverse Bregman $\mathcal{B}_A(\theta^*; \hat{\theta}_n)$? $(\hat{\theta}_n)$ is it for the last iterate or an average ? None of these analysis check all the boxes needed to address $(\star)$.

Boundedness	$-\log$	$\gamma_n \sim \frac{1}{n}$	f	$\hat{\theta}_n$
Variance on Θ (5.31)	✗	✓	✓	✗
Variance at θ^* (5.33)	✗	✓	✗	✓
Optimization gap (5.35)	✓	✗	✗	✓

non-smooth (online) optimization, under bounded (or Lipschitz) gradient assumption on the objective f and strong convexity assumption on the potential A (Bubeck et al., 2015, Th. 4.2(MD) & Th. 6.3(SMD)). In our case, these assumptions do not hold. For instance $A = -\log$ is neither smooth nor strongly convex.

Recently, these assumptions have been relaxed to the α-strong convexity and β-smoothness of f *relative* to a reference function A, defined as

$$\alpha \mathcal{B}_A(x; y) \leq \mathcal{B}_f(x; y) \leq \beta \mathcal{B}_A(x; y) \ .$$

When $A = \|\cdot\|^2$, we recover the standard smoothness and gradient descent. These conditions ensure the linear convergence of MD with mirror map ∇A (Birnbaum et al., 2011; Bauschke et al., 2017; Lu et al., 2018), even when f is not smooth, and A not strongly convex.

For exponential families, MAP perfectly fits into this framework, as

$$f(\theta) = A(\theta) - \mathbb{E}\left[\langle T(X), \theta \rangle\right]$$

is 1-smooth and 1-strongly convex relative to A. Our goal is then to find an applicable convergence rate for SMD under relative smoothness.

5.6.2 Bounding the Randomness

To analyze stochastic algorithms, one also needs to quantify the randomness of stochastic gradients $g(\theta)$. While many assumptions exist for SGD (Khaled and Richtárik, 2020, §3 for a modern review), only a few have been adapted to SMD with relative smoothness (Hanzely and Richtárik, 2021; Dragomir et al., 2021; D'Orazio et al., 2021), but they have so far been lacking concrete examples. We review these analyses in the light of the MAP and provide a summary in Table 5.1.

Analogs of the Variance

Let us introduce the symmetrized Bregman induced by A^*, written $\mathcal{S}_{A^*}(\mu_1;\mu_2) = \mathcal{B}_{A^*}(\mu_1;\mu_2)+\mathcal{B}_{A^*}(\mu_2;\mu_1)$. Hanzely and Richtárik (2021) assume that the expectation of $\mathcal{S}$ between stochastic and deterministic updates verifies

$$\mathbb{E}_g\left[\mathcal{S}_{A^*}(\hat{\mu}_n - \gamma g(\hat{\theta}_n); \hat{\mu}_n - \gamma\nabla f(\hat{\theta}_n))\right] \leq \gamma^2 C \tag{5.31}$$

for all possible iterates $\hat{\theta}_n$, relevant step-sizes γ and for some constant C. When $A(\theta) = \frac{1}{2}\|\theta\|^2$, this definition recovers the variance of the stochastic gradient

$$\mathbb{E}_g\left[\|\nabla f(\theta) - g(\theta)\|^2\right] \leq C\ . \tag{5.32}$$

Under this assumption, Hanzely and Richtárik (2021, Lem.4.8) prove a $O(1/n)$ convergence rate on function values with $O(1/n)$ step-sizes and tail averaging (Lacoste-Julien et al., 2012) in primal space Θ.

Dragomir et al. (2021) define the assumption

$$\mathbb{E}_g[\mathcal{B}_{A^*}(\hat{\mu}_n - 2\gamma g(\theta_*), \hat{\mu}_n)] \leq 2\gamma^2 C\ . \tag{5.33}$$

When $A(\theta) = \frac{1}{2}\|\theta\|^2$, we recover the variance of the gradients at the optimum, which is weaker than (5.32),

$$\mathbb{E}_g\left[\|\nabla g(\theta^*)\|^2\right] \leq C\ . \tag{5.34}$$

Using their descent lemma (Dragomir et al., 2021, Eq. (12)) with the $O(1/n)$ step-size used by Gower et al. (2019, Th. 3.2) for SGD, we obtain a $O(1/n)$ convergence rate, on the Bregman with *reversed* arguments $\mathcal{B}_A(\theta^*;\hat{\theta}_n)$.

These two analyses seem promising for ($\star$), but none of these assumptions hold in front of barrier objectives such as the $-\log$ from §5.4.1. Indeed, they both assume their bound holds uniformly for every possible iterate $\hat{\theta}_n$. Yet $\mathcal{N}(0,\sigma^2)$ has a positive mass around 0. This means that $\hat{\mu}_n$ can get arbitrarily close from 0, where the $-\log$ is unbounded, along with the associated Bregman divergences (5.31) and (5.33). In general, this uniform bound over $\hat{\mu}_n$ cannot hold for *barrier* objectives – functions exploding to infinity in some finite point of space.

Both of their proofs hold if we add an expectation over $\hat{\mu}_n$ to their assumption. However, this is not helpful, as verifying the assumption becomes as hard as the initial problem. For instance, the expectation of (5.31) over $\hat{\mu}_n$ is an upper bound on the variance term of (5.30) (cf App. 5.E). Confronted with this difficulty, we investigate an alternative definition of variance.

Bounded Optimality Gap

Inspired by Loizou et al. (2021), D'Orazio et al. (2021) explore the hypothesis

$$\min_\theta f(\theta) - \mathbb{E}_X\left[\min_\theta f_X(\theta)\right] \leq C\ , \tag{5.35}$$

where f_X is a stochastic estimate of $f = \mathbb{E}[f_X]$. In our case $f_X(\theta) = -\log p(X \,|\, \theta)$. In other words, this lower bounds the expectation of the minimum of the stochastic estimates. For probabilistic models, such a bound is finite as soon as the model cannot give infinite density to any data point x. This holds, for instance, for discrete distributions because the probability mass is upper bounded by 1; however, it rules out many families. In the case of normal distributions $\mathcal{N}(m, \sigma^2)$, setting $m = x$ and $\sigma^2 \to 0$ gets $p_\theta(x) \to +\infty$,. We have a similar behavior for gamma distribution with $\alpha = \beta x$ and $\beta \to +\infty$, or with the beta distribution with $\alpha = \beta \frac{x}{1-x}$ and $\beta \to +\infty$. Other counter-examples include inverse Gaussians, log-normal, gamma, inverse gamma.

It is possible to overcome this limitation by treating batches of samples as single samples by averaging sufficient statistics, e.g.,. $Y = \{X_1, \dots, X_k\}$ and $T(Y) = \frac{1}{k} \sum_i T(X_i)$. For instance, a multivariate normal of dimension d cannot attribute infinite density to $d + 1$ samples that are not in an affine subspace.

Overall, (5.35) can partially handle barrier objectives, but it fails to account for the step-size $\gamma_n = \frac{1}{n_0+n}$, as D'Orazio et al. (2021, Thm.1) only proves linear convergence to a variance ball of size $\frac{C}{\alpha}$ under constant step-size. This is in contrast with Dragomir et al. (2021) which can handle decreasing step-sizes but not barrier objectives. Proving convergence of stochastic mirror descent on barrier loss remains an open problem.

5.7 Conclusion

Despite the MLE and MAP estimators in the exponential family being classical and known in statistics for decades, we highlighted in this paper open problems to bound their frequentist risk (the expected KL) in a non-asymptotic way. We reviewed some partial results, such as a large sample analysis that describes how many samples are needed to ensure a locally quadratic regime (Kakade et al., 2010; Ostrovskii and Bach, 2021) for which rates are known. We also related this problem to the one of obtaining convergence rates in stochastic optimization, observing that MAP fits the framework of stochastic mirror descent with relative smoothness assumptions. Nevertheless, none of the current analyses of SMD hold for the MAP, even on a simple family such as $\mathcal{N}(0, \sigma^2)$, thus revealing an area for progress in non-Euclidean optimization. In writing this paper, we hope to attract attention to this fundamental problem, leading to progress in both optimization and statistics.

5.A Proofs for Gaussian Variance

In this section, we prove the results mentioned in §5.4.1, and add some context and experimental observations. As mentioned in the main text, the centered gaussian $\mathcal{N}(0, \sigma^2)$ has sufficient statistic $T(X) = X^2$ which follows a gamma distribution $\Gamma(\frac{1}{2}, \frac{1}{2\sigma^2})$.

In general, if X is part of the exponential family, then $T(X)$ is part of the natural exponential family with the appropriate support and base measure, with the same log-partition function as X up to constants. MLE and MAP only depend on $T(X)$, not X, so their performance only depends on the distribution of $T(X)$.

In this section we derive results for samples from a general gamma distribution $X \sim \Gamma(\alpha, \beta)$ with known shape parameter α, but unknown rate parameter β. Results for the Gaussian follow by taking $\alpha = \frac{1}{2}$, We also immediately get results for exponential distributions by taking $\alpha = 1$. For instance for the MLE we derive the following theorem:

Theorem 5.A.1 (MLE Upper Bound)**.** *Consider an exponential family such that $T(X)$ is a gamma $\Gamma(\alpha, \beta)$ with known shape α. the expected KL between μ_* and the MLE $\hat{\mu}_n$ is infinite when $\alpha n \leq 1$ and otherwise upper bounded by*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n)\right] \leq \frac{1}{2n} + \frac{1}{n(n\alpha - 1)} \,. \tag{5.36}$$

To obtain the result for Gaussian variance (see Theorem 5.4.1), it suffices to set $\alpha = \frac{1}{2}$ in Theorem 5.A.1.

In this section, we review useful properties of the gamma distribution and associated Bregman divergence in §5.A.1. Then we prove theorem 5.A.1 in §5.A.2. Then we prove an extension of theorem 5.4.1 in §5.A.3, and prove a useful lemma about the expectation of the natural parameter of the MAP in §5.A.4, in order to prove upper bounds for the MAP in §5.A.5. Finally we numerically investigate the effect of prior hyper-parameters in §5.A.6.

5.A.1 Gamma Distribution

The density of $\Gamma(\alpha, \beta)$ reads

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \,. \tag{5.37}$$

When α is known, it can be cast as an exponential family (5.5) with sufficient statistic $T(x) = x$, domain $\mathcal{X} = \mathbb{R}_+$ and base measure $\nu(x) \propto x^{\alpha-1}$. Then the natural parameter is $\theta = -\beta < 0$ and the log-partition function is

$$A(\theta) = -\alpha \log(-\theta) + \log \Gamma(\alpha) \,. \tag{5.38}$$

From there we find that the mean parameter is $\mu = \frac{\alpha}{-\theta} > 0$ and the entropy has the same form as the log-partition $A^*(\mu) = -\alpha \log(\mu) + \text{cst}$. This means that the primal and dual Bregman divergences have the same form as well

$$\mathcal{B}_{A^*}(\mu_*; \mu_n) = \alpha \left(\frac{\mu_*}{\mu_n} - 1 - \log \frac{\mu_*}{\mu_n} \right) = \alpha\phi(\frac{\mu_*}{\mu_n}), \tag{5.39}$$

$$\mathcal{B}_{A}(\theta_n; \theta_*) = \alpha \left(\frac{\theta_n}{\theta_*} - 1 - \log \frac{\theta_n}{\theta_*} \right) = \alpha\phi(\frac{\theta_n}{\theta_*}), \tag{5.40}$$

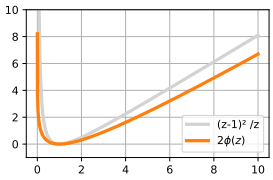


Figure 5.4 – Illustration of $\phi(z)$ (orange) and its upper bound $\phi(z) + \phi(z^{-1})$ (grey). They both are barriers near 0.

where these 2 lines are equal, and ϕ measures the discrepancy between the ratio $\frac{\theta_n}{\theta_*} = \frac{\mu_*}{\mu_n}$ and 1 via the function

$$\phi(z) := z - 1 - \log(z), \tag{5.41}$$

illustrated in orange in Figure 5.4. To derive the upper bound for the MAP, due to the difficulty of finding a closed form for the expectation of the logarithm, we focus on the symmetrized Bregman instead

$$\mathcal{S}_{A^*}(\mu_*, \mu_n) := \mathcal{B}_{A^*}(\mu_*; \mu_n) + \mathcal{B}_{A^*}(\mu_n; \mu_*) \tag{5.42}$$

$$= \alpha\phi(\frac{\mu_*}{\mu_n}) + \alpha\phi(\frac{\mu_n}{\mu_*}) \tag{5.43}$$

$$= \alpha(\frac{\mu_*}{\mu_n} - 1 + \frac{\mu_n}{\mu_*} - 1) \ , \tag{5.44}$$

which verifies $\mathcal{B}_{A^*}(\mu_*; \mu_n) \leq \mathcal{S}_{A^*}(\mu_*, \mu_n)$. Writing $z = \frac{\mu_*}{\mu_n}$ this is equivalent to

$$\phi(z) \leq \phi(z) + \phi(z^{-1}) = z - 1 + \frac{1}{z} - 1 = \frac{(z-1)^2}{z} \ ,$$

which is illustrated by the grey upper bound in Figure 5.4.

5.A.2 Proof for the MLE

Proof. Since $T(X)$ follows a gamma distribution $\Gamma(\alpha, \beta)$, the MLE is a scaled sum of gammas $\hat{\mu}_n = \frac{1}{n}\sum_i T(X_i)$. As such it is also a gamma with parameter $\Gamma(n\alpha, n\beta)$ and expectation $\frac{n\alpha}{n\beta} = \frac{\alpha}{\beta} = \mu_*$. If we consider the ratio $\frac{\hat{\mu}_n}{\mu_*}$, it is also a gamma with parameter $\Gamma(n\alpha, n\alpha)$. Its inverse follows an inverse gamma distribution with expectation

$$\mathbb{E}\left[\frac{\mu_*}{\hat{\mu}_n}\right] = \begin{cases} \frac{n\alpha}{n\alpha - 1} & \text{if } n\alpha > 1, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.45}$$

which implies that for $n\alpha > 1$,

$$\mathbb{E}\left[\frac{\mu_*}{\hat{\mu}_n}\right] - 1 = \frac{n\alpha}{n\alpha - 1} - 1 = \frac{1}{n\alpha - 1} \tag{5.46}$$

There is also a closed form solution for the expected logarithm of a gamma. Indeed, the sufficient statistic of a gamma is $(X, \log(X))$, so one can apply formula (5.7) on the log-partition of a gamma to get

$$\mathbb{E}\left[\log \frac{\hat{\mu}_n}{\mu_*}\right] = \psi(n\alpha) - \log(n\alpha), \tag{5.47}$$

where ψ is the digamma function. Consequently the suboptimality of the MLE has a closed form solution

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n)\right] = \alpha\mathbb{E}\left[\frac{\mu_*}{\hat{\mu}_n} - 1 + \log\left(\frac{\hat{\mu}_n}{\mu_*}\right)\right], \tag{5.48}$$

$$= \begin{cases} \alpha(\frac{1}{n\alpha-1} + \psi(n\alpha) - \log(n\alpha)), & \text{if } n\alpha > 1, \\ +\infty \quad \text{otherwise.} \end{cases} \tag{5.49}$$

Surprisingly, for a gaussian variance where $\alpha = \frac{1}{2}$, we need 3 samples or more for the expected loss to be bounded. When the expectation is finite, we can get a more interpretable formula using known bounds on the digamma function,

$$-\frac{1}{x} \leq \psi(x) - \log(x) \leq -\frac{1}{2x} = -\frac{1}{x} + \frac{1}{2x}\ , \tag{5.50}$$

giving, for $n\alpha > 1$,

$$\frac{\alpha}{n\alpha - 1} - \frac{\alpha}{n\alpha} \leq \mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n)\right] \leq \frac{\alpha}{n\alpha - 1} - \frac{\alpha}{n\alpha} + \frac{\alpha}{2n\alpha} \tag{5.51}$$

$$\iff \frac{1}{n(n\alpha - 1)} \leq \mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n)\right] \leq \frac{1}{2n} + \frac{1}{n(n\alpha - 1)}\ , \tag{5.52}$$

so we get a $\Omega(n^{-2})$ lower bound and a $O(n^{-1}) + O(n^{-2})$ upper bound. □

5.A.3 Multivariate MLE

For the sake of simplicity, in higher dimension we focus on Gaussian covariance estimation and avoid the general Wishart discussion. In higher dimensions, $X \sim \mathcal{N}(0, \mu_*)$, $X \in \mathbb{R}^d$, $T(X) = XX^\top$, and the mean parameter μ_* is a $d \times d$ symmetric, positive definite covariance matrix with $p = \frac{d(d+1)}{2}$ degrees of freedom. Note that here d denotes the dimensionality of the data X, rather than the dimensionality of the parameters μ_*.

Theorem 5.A.2 (Multivariate MLE Upper Bound)**.** *The MLE of the covariance matrix of $X_i \sim \mathcal{N}(0, \mu_*)$ is $\hat{\mu}_n = \frac{1}{n}\sum_i X_i X_i^\top \in \mathbb{R}^{d\times d}$ with $p = \frac{d(d+1)}{2}$ degrees of freedom. The expected KL divergence between μ_* and $\hat{\mu}_n$ is infinite when $n \leq d+1$ and otherwise upper bounded by*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n^{MLE(d)})\right] \leq \frac{p}{2n} + \frac{p(d+2)}{n(n-d-1)}\ , \forall n > d+1\ . \tag{5.53}$$

We see from (5.23) that this bound is asymptotically tight.

Proof. The entropy of X is a negative log-determinant $A^*(\mu) = -\log\det(\mu)$, whose gradient is the negative matrix inverse $\nabla A^*(\mu) = -\mu^{-1}$. The associated Bregman divergence is

$$\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n) = \frac{1}{2}(\operatorname{Tr}(\mu_*\hat{\mu}_n^{-1}) - d - \log\det(\mu_*\hat{\mu}_n^{-1})) \; . \tag{5.54}$$

Thanks to the linearity of the trace, the expectation becomes

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n)\right] = \frac{1}{2}(\operatorname{Tr}(\mathbb{E}\left[\mu_*\hat{\mu}_n^{-1}\right]) - d + \mathbb{E}\left[\log\det(\hat{\mu}_n\mu_*^{-1}))\right]) \; . \tag{5.55}$$

When the estimator is the MLE, $\hat{\mu}_n = \frac{1}{n}\sum_i X_i X_i^\top$, then we define the mean parameter "ratio" as

$$\boldsymbol{V} := n\mu_*^{-\frac{1}{2}}\hat{\mu}_n\mu_*^{-\frac{1}{2}} = \sum_i (\mu_*^{-\frac{1}{2}}X_i)(\mu_*^{-\frac{1}{2}}X_i)^\top,$$

such that $\operatorname{Tr}(\mu_*\hat{\mu}_n^{-1}) = n\operatorname{Tr}(\boldsymbol{V}^{-1})$. But $\mu_*^{-\frac{1}{2}}X_i \sim \mathcal{N}(0, \boldsymbol{I})$, so that $\boldsymbol{V}$ is sampled from a Wishart $\mathcal{W}(\boldsymbol{I}, n)$, where $\boldsymbol{I}$ stands for the identity matrix of order d. Recall that $\mathbb{E}[\boldsymbol{V}] = n\boldsymbol{I}$. Thanks to $\log\det$ being a sufficient statistic of the Wishart, and the natural to mean parameter formula (5.7), we have a closed form for the expected log-determinant of a Wishart

$$\mathbb{E}[\log\det\boldsymbol{V}] = \psi_d(\frac{n}{2}) + d\log 2 \; ,$$

where $\psi_d(\frac{n}{2}) = \sum_{i=0}^{d-1}\psi(\frac{n-i}{2})$ is the multivariate digamma function. The expectation of an inverse Wishart is straightforward to compute from the density and the log-partition function

$$\mathbb{E}[\boldsymbol{V}^{-1}] = \begin{cases} +\infty \text{ if } n \leq d+1 \\ \frac{\boldsymbol{I}}{n-d-1} \text{ otherwise,} \end{cases} \tag{5.56}$$

which proves the infinite part of the statement. Consider now the case $n > d+1$. Using

$$\operatorname{Tr}(\mathbb{E}\left[\mu_*\hat{\mu}_n^{-1}\right]) - d = n\operatorname{Tr}(\mathbb{E}[\boldsymbol{V}^{-1}]) - d = \frac{nd}{n-d-1} - d = \frac{d(d+1)}{n-d-1} = \frac{2p}{n-d-1}, \tag{5.57}$$

and putting it all together, we get the following closed form for the expectation of the divergence

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n^{\mathrm{MLE(d)}})\right] = \frac{p}{n-d-1} + \frac{1}{2}\left(\psi_d(\frac{n}{2}) - d\log(\frac{n}{2})\right). \tag{5.58}$$

To bound ψ_d, we can use the same bound as in the univariate case, $\psi(x) \leq \log(x) - \frac{1}{2x}$, to get

$$\psi_d(\frac{n}{2}) - d\log(\frac{n}{2}) = \sum_{i=0}^{d-1}(\psi(\frac{n-i}{2}) - \log(\frac{n}{2})) \tag{5.59}$$

$$\leq \sum_{i=0}^{d-1}(\log(\frac{n-i}{2}) - \frac{1}{n-i} - \log(\frac{n}{2})) \tag{5.60}$$

$$= \sum_{i=0}^{d-1}(\log(1-\frac{i}{n}) - \frac{1}{n-i}) \,. \tag{5.61}$$

We can bound sum of reciprocals $\sum_{i=0}^{d-1} \frac{1}{n-i}$ by the typical bound on the harmonic sum $H_n = \sum_{k=1}^{n} \frac{1}{k}$,

$$\frac{1}{2n+1} \leq H_n - \log(n) - \gamma \leq \frac{1}{2n-1},$$

where γ is the Euler constant. Then

$$-\sum_{i=0}^{d-1}\frac{1}{n-i} = H_{n-d} - H_n \leq \log(n-d) + \gamma + \frac{1}{2(n-d)-1} - \log(n) - \gamma - \frac{1}{2n+1} \tag{5.62}$$

$$= \log(1-\frac{d}{n}) + \frac{2(d+1)}{(2n+1)(2(n-d)-1)} \tag{5.63}$$

$$< \log(1-\frac{d}{n}) + \frac{d+1}{2n(n-d-1)} \,. \tag{5.64}$$

Plugging this back into (5.61) yields

$$\psi_d(\frac{n}{2}) - d\log(\frac{n}{2}) \leq \sum_{i=0}^{d}\log(1-\frac{i}{n}) + \frac{d+1}{2n(n-d-1)} \,, \tag{5.65}$$

where the sum now goes up to d. To bound the remaining sum, we use $\log(1+x) \leq x$ (for $x > -1$) to get

$$\sum_{i=0}^{d}\log(1-\frac{i}{n}) \leq \sum_{i=0}^{d} -\frac{i}{n} = -\frac{d(d+1)}{2n} = -\frac{p}{n} \,. \tag{5.66}$$

Putting those bounds together in Eq. (5.58) and reorganizing yields

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n^{\mathrm{MLE(d)}})\right] < p(\frac{1}{n-d-1} - \frac{1}{2n}) + \frac{d+1}{4n(n-d-1)} \tag{5.67}$$

$$= p\frac{n+d+1}{2n(n-d-1)} + \frac{d(d+1)/2}{2dn(n-d-1)} \tag{5.68}$$

$$= p\frac{n-d-1}{2n(n-d-1)} + p\frac{2(d+1)}{2n(n-d-1)} + \frac{p}{2dn(n-d-1)} \tag{5.69}$$

$$= \frac{p}{2n} + \frac{p(d+1+\frac{1}{2d})}{n(n-d-1)} \tag{5.70}$$

$$\leq \frac{p}{2n} + \frac{p(d+2)}{n(n-d-1)} \ , \forall n > d+1 \ . \tag{5.71}$$

where $p = \frac{d(d+1)}{2}$ and the last inequality used $\frac{1}{2d} \leq 1$. □

5.A.4 Bounding the Expected Natural Parameter for the MAP

Before proving a convergence rate for the MAP, we need to bound the expectation of its inverse, hence the following lemma. We introduce the notation $(z)_+ = \max(0, z)$.

Lemma 5.A.3 (Expected MAP natural parameter)**.** *Define the variable $a = n_0\frac{\mu_0}{\mu^*}$, which characterizes the importance of the prior relative to the true parameters. The expectation of the natural parameter of a MAP of $\Gamma(\alpha, \beta)$ is bounded by*

$$\frac{n_0+n}{a+n} = \frac{\mu^*}{\mu_n} \leq \mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] = \mathbb{E}\left[\frac{\hat{\theta}_n}{\theta^*}\right] \leq \frac{n_0+n}{a+(n-\frac{1}{\alpha})_+} \ , \forall n \geq 0 \ . \tag{5.72}$$

Proof. The lower bound can be readily obtained by applying Jensen's inequality to the convex function $x \mapsto \frac{1}{x}$ for $x > 0$. The upper bound requires more work. To start, let us plug in the definition of $\hat{\mu}_n$

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] = \mathbb{E}\left[\frac{\hat{\theta}_n}{\theta^*}\right] = \mathbb{E}\left[\frac{(n_0+n)\mu^*}{n_0\mu_0 + \sum_i X_i}\right] = \mathbb{E}\left[\frac{n_0+n}{n_0\frac{\mu_0}{\mu^*} + \sum_i \frac{X_i}{\mu^*}}\right] = \mathbb{E}\left[\frac{n_0+n}{a+\Gamma(n\alpha,\alpha)}\right] \ , \tag{5.73}$$

where $\sum_i \frac{X_i^2}{\mu^*} \sim \Gamma(n\alpha, \alpha)$ is a gamma random variable and $a = n_0\frac{\mu_0}{\mu^*}$. Further note that

$$\mathbb{E}\left[\frac{n_0+n}{a+\Gamma(n\alpha,\alpha)}\right] = \frac{n_0+n}{a}\mathbb{E}\left[\frac{1}{1+\Gamma(n\alpha,a\alpha)}\right] \ . \tag{5.74}$$

This kind of integrals can be expressed with generalized exponential integral functions

$$E_k(z) = \int_1^\infty \frac{e^{-zt}}{t^k} dt \ , \tag{5.75}$$

with the formula (Olver et al., 2021, Eq. 8.19.4)

$$\mathbb{E}\left[\frac{1}{1+\Gamma(\alpha,\beta)}\right] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{x^{\alpha-1}e^{-\beta x}}{1+x} dx = \beta e^\beta E_\alpha(\beta) \ . \tag{5.76}$$

Overall we get

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] = (n_0+n)\alpha e^{a\alpha} E_{n\alpha}(a\alpha) \tag{5.77}$$

Now our goal is to bound this generalized exponential integral with simpler functions. Fortunately, mathematicians have been working on these integrals for decades. For instance , we have the general bound (Olver et al., 2021, Eq. 8.19.21)

$$e^x E_k(x) \leq \frac{1}{x+k-1}, \ \forall k > 1 \tag{5.78}$$

$$\iff \mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] \leq \frac{(n_0+n)\alpha}{a\alpha+n\alpha-1} = \frac{n_0+n}{a+n-\frac{1}{\alpha}}, \ \forall n > \frac{1}{\alpha} \ . \tag{5.79}$$

We are left with a special case when $n \leq \alpha$. Then we can use the trivial bound

$$\mathbb{E}\left[\frac{1}{a+X^2}\right] < \frac{1}{a}. \tag{5.80}$$

to conclude the proof. □

When $n < \frac{1}{\alpha}$, it is possible to get a much tighter bound by exploiting the recurrence relationship (Olver et al., 2021, Eq. 8.19.12)

$$\alpha E_{\alpha+1}(\beta) + \beta E_\alpha(\beta) = e^{-\beta} \tag{5.81}$$

and combining it with the inequality (Olver et al., 2021, Eq. 8.19.21) to get

$$\beta e^\beta E_\alpha(\beta) = 1 - \alpha e^\beta E_{\alpha+1}(\beta) \leq 1 - \frac{\alpha}{\alpha+1+\beta} = \frac{\beta+1}{\alpha+\beta+1} \ . \tag{5.82}$$

Plugging this inequality back into (5.76), we get the following upper bound for the MAP:

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] \leq \frac{n_0+n}{a} \frac{a\alpha+1}{n\alpha+a\alpha+1} = \frac{n_0+n}{a+n+\frac{1}{\alpha}}(1+\frac{1}{a\alpha}) \ . \tag{5.83}$$

Unfortunately, this formula does not yield an elegant convergence rate, so we keep it out of the lemma.

5.A.5 Proof of MAP Bound

We did not find a closed form or an upper bound for the expected logarithm of the MAP $\mathbb{E}\left[\log \frac{\hat{\mu}_n}{\mu^*}\right]$. Consequently, we derived a bound for the symmetrized Bregman (5.44) instead. This bound is asymptotically tight for the Bregman, up to a factor 2.

There are several ways to write down the convergence rate. The Gaussian variance example can be written in a particularly simple form, so we give it a special treatment in §5.A.5, corresponding to the theorem displayed in the main text. We make a more general statement about gamma distributions in §5.A.5.

Gaussian Variance

Theorem 5.A.4 (MAP Bound)**.** *The expected symmetrized Bregman (5.44) of the MAP of $\mathcal{N}(0, \mu^*)$ with prior hyper-parameters (n_0, μ_0) is upper bounded as*

$$\mathbb{E}\left[\mathcal{S}_{A^*}(\mu^*; \hat{\mu}_n^{\text{MAP}})\right] \leq \begin{cases} \mathcal{S}_{A^*}(\mu_*; \mu_0) & \text{if } n = 0, \\ \frac{1}{2(n_0+1)} + b_1 & \text{if } n = 1, \\ \frac{1}{n_0 \frac{\mu_0}{\mu^*} + n - 2} + b_n & \text{if } n \geq 2, \end{cases} \qquad \text{where } b_n = \frac{(1 + \frac{1}{n_0} - \frac{\mu_0}{\mu^*})^2}{2(\frac{\mu_0}{\mu^*} + \frac{(n-2)_+}{n_0})(1 + \frac{n}{n_0})} \in O(\tfrac{1}{n}) \tag{5.84}$$

Proof. When $n = 0$, the inequality is an equality. For $n > 0$, we expand the symmetrized Bregman (5.44) with $\alpha = \frac{1}{2}$ to get

$$\mathbb{E}\left[\mathcal{S}_{A^*}(\mu^*; \hat{\mu}_n)\right] \leq \frac{1}{2}(\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 + \mathbb{E}\left[\frac{\hat{\mu}_n}{\mu^*}\right] - 1) \,. \tag{5.85}$$

The expectation of $\hat{\mu}_n$ is straightforward

$$\mathbb{E}\left[\frac{\hat{\mu}_n}{\mu^*}\right] - 1 = \frac{n_0\mu_0 + n\mu^*}{(n_0 + n)\mu^*} - 1 = \frac{a + n}{n_0 + n} - 1 = \frac{a - n_0}{n_0 + n} \quad \text{where} \quad a := n_0\frac{\mu_0}{\mu^*} \,. \tag{5.86}$$

There remains the more problematic term with the expectation of the inverse mean parameter, for which we use the bound derived in Lemma 5.A.3.

When $n \geq 2$, we get

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 \leq \frac{n_0 + n}{a + n - 2} - 1 = \frac{n_0 - a + 2}{a + n - 2} = \frac{2}{a + n - 2} + \frac{n_0 - a}{a + n - 2} \tag{5.87}$$

so putting it all together we get

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*;\hat{\mu}_n)\right] \leq \frac{1}{a+n-2} + \frac{n_0-a}{2}\left(\frac{1}{a+n-2} - \frac{1}{n_0+n}\right) \tag{5.88}$$

$$= \frac{1}{a+n-2} + \frac{(n_0-a+1)-1}{2}\frac{(n_0-a+1)+1}{(a+n-2)(n_0+n)} \tag{5.89}$$

$$= \frac{1}{a+n-2} + \frac{(n_0-a+1)^2-1}{2(a+n-2)(n_0+n)} \tag{5.90}$$

$$\leq \frac{1}{a+n-2} + \frac{(n_0-a+1)^2}{2(a+n-2)(n_0+n)} \tag{5.91}$$

$$= \frac{1}{a+n-2} + \frac{(1+\frac{1}{n_0}-\frac{\mu_0}{\mu^*})^2}{2(\frac{\mu_0}{\mu^*}+\frac{n-2}{n_0})(1+\frac{n}{n_0})}\,. \tag{5.92}$$

When $n=1$ the bound (5.72) on the expected natural parameter gives

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 \leq \frac{n_0+1}{a} - 1 = \frac{n_0+1-a}{a} \tag{5.93}$$

so putting it all together we get

$$2\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*;\hat{\mu}_1)\right] \leq \frac{a-n_0\pm 1}{n_0+1} + \frac{n_0+1-a}{a} \tag{5.94}$$

$$= \frac{1}{n_0+1} + (a-n_0-1)(\frac{1}{n_0+1} - \frac{1}{a}) \tag{5.95}$$

$$= \frac{1}{n_0+1} + \frac{(n_0+1-a)^2}{a(n_0+1)} \tag{5.96}$$

$$= \frac{1}{n_0+1} + \frac{(1+\frac{1}{n_0}-\frac{\mu_0}{\mu^*})^2}{\frac{\mu_0}{\mu^*}(1+\frac{1}{n_0})} \tag{5.97}$$

so we recover the same bias term as when $n \geq 2$. □

Gamma with Known Shape

Theorem 5.A.5 (MAP Bound)**.** *Consider an exponential distribution with sufficient statistics coming from a gamma distribution $\Gamma(\alpha,\beta^*)$ with mean parameter $\mu^* = \frac{\alpha}{\beta^*}$. The expected symmetrized Bregman (5.44) of the MAP with prior hyper-parameters (n_0,μ_0) is upper bounded as*

$$\forall n \geq \frac{1}{\alpha}, \qquad \mathbb{E}\left[\mathcal{S}_{A^*}(\mu^*,\hat{\mu}_n)\right] \leq \frac{1}{n_0+n} + \frac{\alpha(\frac{\mu_0}{\mu^*}-\frac{1}{\alpha n_0}-1)^2}{(1+\frac{n}{n_0})(\frac{\mu_0}{\mu^*}+\frac{n-\frac{1}{\alpha}}{n_0})} \tag{5.98}$$

Note that the second term vanishes when $\mu_0 = \mu^*(1 + \frac{1}{\alpha n_0})$. Expressions for $n\alpha < 1$ are less elegant as shown below:

$$\mathbb{E}\left[\mathcal{S}_{A^*}(\mu^*, \hat{\mu}_n)\right] \leq \frac{n_0 + n}{a + n + \frac{1}{\alpha}}(1 + \frac{1}{a\alpha}) - 1 + \frac{a - n_0}{n_0 + n} \quad \text{where} \quad a := n_0\frac{\mu_0}{\mu^*} \,. \tag{5.99}$$

Proof. We expand the symmetrized Bregman (5.44) to get

$$\mathbb{E}\left[\mathcal{S}_{A^*}(\mu^*; \hat{\mu}_n)\right] \leq \alpha(\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 + \mathbb{E}\left[\frac{\hat{\mu}_n}{\mu^*}\right] - 1) \,. \tag{5.100}$$

The expectation of $\hat{\mu}_n$ is straightforward

$$\mathbb{E}\left[\frac{\hat{\mu}_n}{\mu^*}\right] - 1 = \frac{n_0\mu_0 + n\mu^*}{(n_0 + n)\mu^*} - 1 = \frac{a + n}{n_0 + n} - 1 = \frac{a - n_0}{n_0 + n} \quad \text{where} \quad a := n_0\frac{\mu_0}{\mu^*} \,. \tag{5.101}$$

There remains the more problematic term with the expectation of the inverse mean parameter, for which we use the bound derived in Lemma 5.A.3 when $n\alpha \geq 1$

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 \leq \frac{n_0 + n}{a + n - \frac{1}{\alpha}} - 1 = \frac{n_0 - a + \frac{1}{\alpha}}{a + n - \frac{1}{\alpha}} \tag{5.102}$$

so putting it all together we get

$$\begin{aligned}
\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n)\right] &\leq \alpha(\frac{n_0 - a + \frac{1}{\alpha}}{a + n - \frac{1}{\alpha}} + \frac{a - n_0 \pm \frac{1}{\alpha}}{n_0 + n}) && (5.103)\\
&= \alpha(n_0 + \frac{1}{\alpha} - a)\left(\frac{1}{a + n - \frac{1}{\alpha}} - \frac{1}{n_0 + n}\right) + \frac{1}{n_0 + n} && (5.104)\\
&= \frac{1}{n_0 + n} + \alpha\frac{(n_0 + \frac{1}{\alpha} - a)^2}{(n_0 + n)(a + n - \frac{1}{\alpha})} && (5.105)\\
&= \frac{1}{n_0 + n} + \alpha\frac{(1 + \frac{1}{\alpha n_0} - \frac{\mu_0}{\mu^*})^2}{(1 + \frac{n}{n_0})(\frac{\mu_0}{\mu^*} - \frac{1}{\alpha n_0} + \frac{n}{n_0})} \,. && (5.106)
\end{aligned}$$

□

5.A.6 On the Choice of a Prior

The optimal μ_0 is larger than μ^* for small n_0 and small n. Indeed, the upper bound (5.19) has a bias term that is 0 when $\frac{mu_0}{\mu^*} = 1 + \frac{1}{n_0}$, e.g. for large values of n_0, it is $\mu_0 = \mu^*$ is the best prior, but for small n_0, one better sets larger values for μ_0. In Figure 5.5, we observe this behavior numerically.

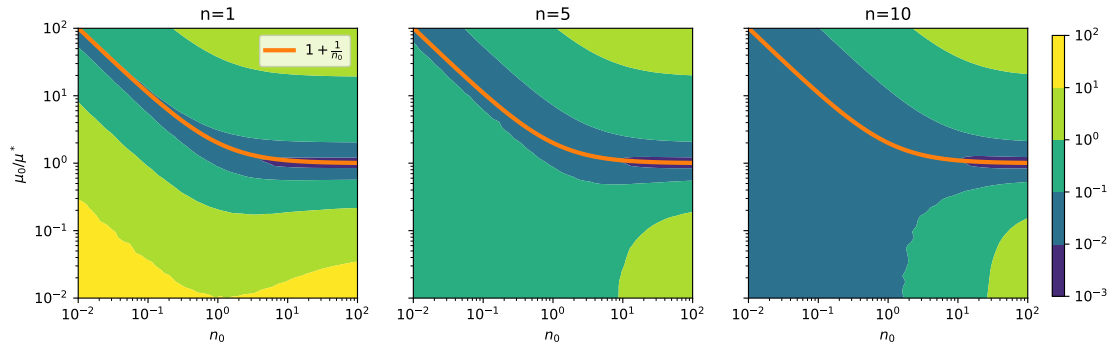


Figure 5.5 – On the optimal priors (n_0, μ_0)**.** Contours of $\mathbb{E}\left[D_{\mathrm{KL}}(\theta^*, \hat{\theta}_n)\right]$ for the Gaussian variance with $n \in \{1, 5, 10\}$, $\mu^* = 1$ and n_0, μ_0 spanning $[10^{-2}, 10^2]$. The expectation was estimated with 10^4 draws for each value of n_0, μ_0 We observe that the line $\frac{\mu_0}{\mu^*} = 1 + \frac{1}{n_0}$ coincides with the bottom valley of this landscape.

5.B Complements on Gaussians

In this section, we prove that for a Gaussian, the entropy and log-partition functions are self-concordant. We also provide complementary illustrations of these functions in Fig. 5.6, their gradients (e.g. the mirror maps) in Fig. 5.7, and paths taken by MLE and MAP in Fig. 5.8 .

The Gaussian log-partition function and entropy are, up to constants,

$$A(\theta) = \frac{\theta_1^2}{-4\theta_2} - \frac{1}{2}\log(-\theta_2) \tag{5.107}$$

$$A^*(\mu) = -\frac{1}{2}\log(\mu_2 - \mu_1^2)\ , \tag{5.108}$$

where $\theta_1 \in \mathbb{R}, \theta_2 < 0$ and $\mu_1 \in \mathbb{R}, \mu_2 > \mu_1^2$. We provide definitions of self-concordance in Section 5.D.

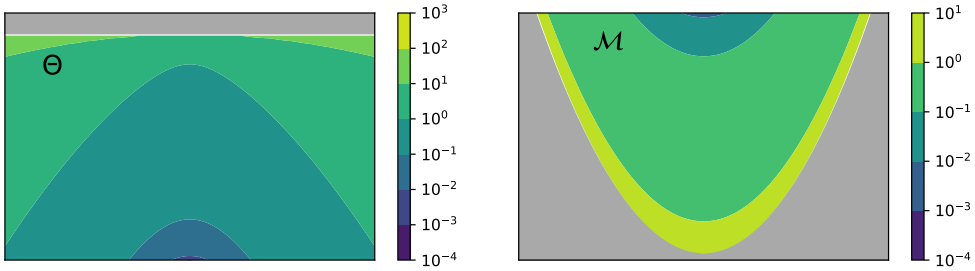


Figure 5.6 – (left) Contours of the log-partition function (5.107). (right) Contours of the entropy (5.108).

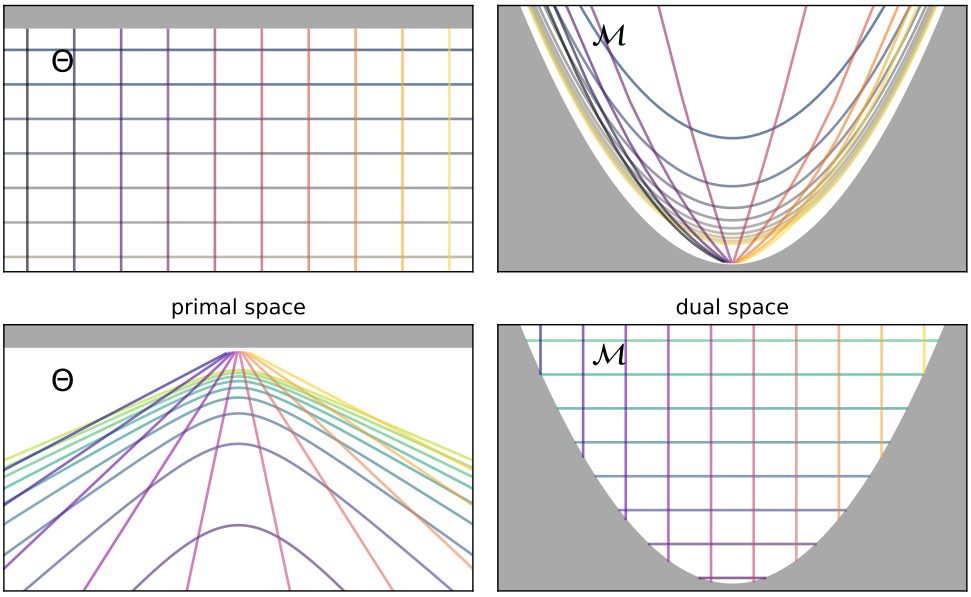


Figure 5.7 – Visualizations of the Gaussian mirror-map. (top) Grid deformation produce by $\nabla A(\theta)$. (bottom) Grid deformation produced by $\nabla A^*(\mu)$.

Entropy is Self-Concordant. Nesterov (2004c, Example 4.1.1.4, p.177) proves that logarithmic barriers for second order regions are self-concordant, that is functions of the form

$$f(\theta) = -\log(\alpha + \langle a, \theta\rangle - \frac{1}{2}\langle \boldsymbol{A}\theta, \theta\rangle) \text{ on } \left\{\theta \in \mathbb{R}^n \,|\, \alpha + \langle a, \theta\rangle - \frac{1}{2}\langle \boldsymbol{A}\theta, \theta\rangle > 0\right\}. \tag{5.109}$$

The entropy (5.108) fits into this definition with $\boldsymbol{A} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ and $a = (0\ 1)^\top$.

Log-partition is Self-Concordant. As proved in Nesterov and Nemirovskii (1994), self-concordance is preserved by Fenchel conjugacy. Since A^* is self-concordant, A is as well. For a more accessible reference, see also Sun and Tran-Dinh (2019, Prop. 6).

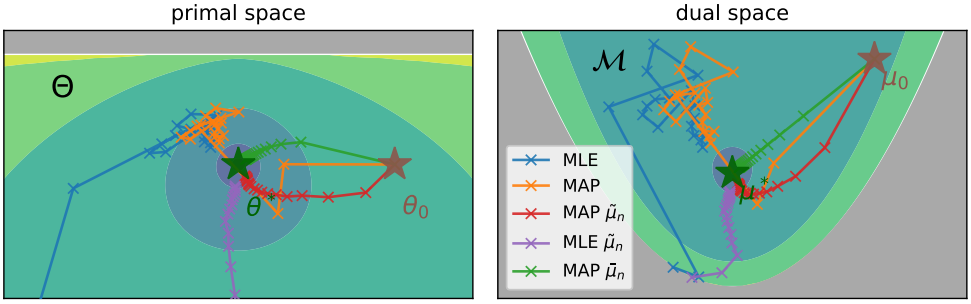


Figure 5.8 – Paths taken by MLE (blue) and MAP (orange) on top of contours for $D_{\mathrm{KL}}(\theta^*||\theta)$. We set $\mu^* = (0,1), \mu_0 = (1,2)$, and $n_0 = 4$, and n varies from 1 to 20. In green, red and purple, we represent the paths respectively taken by the MAP dual expectation $(\bar{\theta}_n, \bar{\mu}_n)$, MAP primal expectation $(\tilde{\theta}_n, \tilde{\mu}_n)$, and MLE primal expectation. Recall that the MLE dual expectation is μ^* itself.

5.C Asymptotic Derivation

In this section we fill-in the lines of §5.5.1 to prove Equation (5.23). Approximating A^* with a second order Taylor expansion yields

$$\mathcal{B}_{A^*}(\mu^*; \mu) = \frac{1}{2}\|\mu - \mu^*\|_{\boldsymbol{F}}^2 + O(\|\mu - \mu^*\|^3),$$

where the norm is induced by the matrix

$$\boldsymbol{F} := \nabla^2 A^*(\mu^*) = \nabla^2 A(\theta^*)^{-1} = \mathrm{Cov}_{\theta^*}(T)^{-1},$$

where the second equality is a general property of convex conjugates. Plugging the MLE (5.2) into this quadratic and expanding it yields

$$\begin{aligned}
\mathbb{E}\,\frac{1}{2}\left\|\frac{1}{n}\sum_i T_i - \mu^*\right\|_{\boldsymbol{F}}^2 &= \frac{1}{2n^2}\sum_i \mathbb{E}\|T_i - \mu^*\|_{\boldsymbol{F}}^2 + \frac{1}{2n^2}\sum_{i\neq j}\mathbb{E}\left[T_i - \mu^*\right]^\top \boldsymbol{F}\,\overbrace{\mathbb{E}\left[T_j - \mu^*\right]}^{0} \\
&= \frac{1}{2n}\,\mathbb{E}\|T_1 - \mu^*\|_{\boldsymbol{F}}^2 \\
&= \frac{1}{2n}\,\mathrm{Tr}(\boldsymbol{F}\;\mathbb{E}\left[(T_1 - \mu^*)(T_1 - \mu^*)^\top\right]) \\
&= \frac{1}{2n}\,\mathrm{Tr}(\boldsymbol{F}\,\mathrm{Cov}_{\theta^*}(T)) = \frac{d}{2n},
\end{aligned}$$

where on the first line we used independence of samples, on the second line we used the fact that samples are identically distributed, and on the third line we used the trace trick along with the linearity of the trace Tr. On the way, this also proves that for the MLE $\|\mu - \mu^*\|^3 \in O(n^{-\frac{3}{2}})$. This yield the final rate for the MLE

$$\mathbb{E}\,\mathcal{B}_{A^*}(\mathbb{E}[T(X)]; \frac{1}{n}\sum_i T_i) = \frac{d}{2n} + O(n^{-\frac{3}{2}})\,. \tag{5.110}$$

For the MAP (5.11), the quadratic decomposes into bias and variance:

$$\mathbb{E}\,\frac{1}{2}\left\|\mu^* - \frac{n_0\mu_0 + \sum_i T(x_i)}{n_0+n}\right\|_{\boldsymbol{F}}^2 = \frac{nd}{2(n+n_0)^2} + \frac{n_0^2}{(n+n_0)^2}\frac{1}{2}\|\mu^*-\mu_0\|_{\boldsymbol{F}}^2 \tag{5.111}$$

$$= \frac{d}{2n} + O\left(\frac{1+\|\mu^*-\mu_0\|_{\boldsymbol{F}}^2}{n^2}\right) . \tag{5.112}$$

This $O(n^{-2})$ term is dominated by the $O(n^{-\frac{3}{2}})$ term from the quadratic approximation of the Bregman, yielding the same first order rate as for the MLE

$$\mathbb{E}\,\mathcal{B}_{A^*}(\mathbb{E}[T(X)]; \frac{n_0\mu_0 + \sum_i T(x_i)}{n_0+n}) = \frac{d}{2n} + O(n^{-\frac{3}{2}}) . \tag{5.113}$$

5.D Self-Concordance

In this section, we define self-concordance and we prove Proposition 5.5.1.

Definition 5.D.1 (Self-concordance)**.** *A convex function is $F : \mathbb{R}^p \to \mathbb{R}$ is self-concordant if it is differentiable 3 times and if for all $w, v \in \mathbb{R}^p$ the function $g(t) = F(w+tv)$ satisfies for all feasible t*

$$|g'''(t)| \leq 2g''(t)^{\frac{3}{2}} . \tag{5.114}$$

Clarification: In the main text in Section 5.5.3, we claimed that the Fenchel conjugate of a 1-dimensional function is also self-concordant. Actually, this is also true in higher dimensions, as proved by Nesterov and Nemirovskii (1994). See Sun and Tran-Dinh (2019, Prop. 6) for a more accessible reference.

5.D.1 Properties of Self-concordant functions

We quickly review some important properties of self-concordant functions, introduced in (Nesterov, 2004c). We start with a some notation. Let A^* be a self-concordant function. Then, we write

- the local norm $\|\cdot\|_x = \sqrt{\langle \nabla^2 A^*(x)\cdot, \cdot\rangle}$
- the distance function $\omega(t) = t - \ln(1+t)$, $t \geq 0$, and its dual $\omega^*(t) = -t - \ln(1-t)$ defined for $t \in [0,1]$.

Note that $\omega^*(t)$ is positive, convex and monotonically increasing for $t \in [0,1]$.

We now present two important results. The first one shows how to convert the local norm $\|y-x\|_y$ using $\|y-x\|_x$.

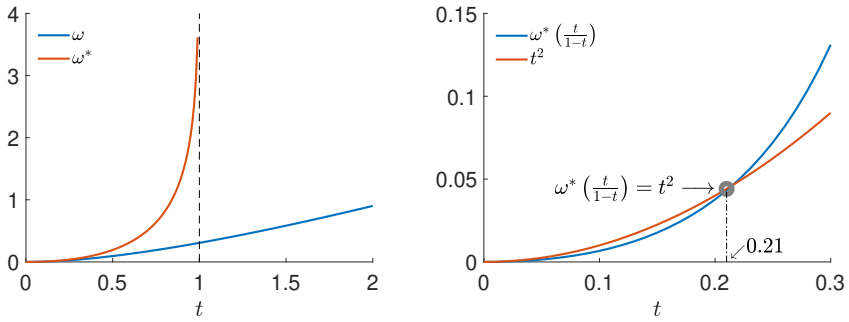


Figure 5.9 – (left) Graph of the distance function ω and its dual ω^*. (right) Graph of the dual of distance function ω^* evaluated at $\frac{t}{1-t}$, compared with t^2. The two curves cross each other at $t \approx 0.21$.

Proposition 5.D.2. *(Conversion of norms, (Nesterov, 2004c, Theorem 4.1.5)) For any x, $y \in \operatorname{Dom} A^*$, if $\|y - x\|_x < 1$, then*

$$\|y - x\|_y \leq \frac{\|y - x\|_x}{1 - \|y - x\|_x}.$$

The next result shows that, if y is sufficiently close to x, then we can bound the Bregman divergence of A^* using the distance function ω^* and local norms.

Proposition 5.D.3. *(Upper bound of self-concordant functions (Nesterov, 2004c, Theorem 4.1.8)) For any x, $y \in \operatorname{Dom} A^*$, if $\|y - x\|_x < 1$, then*

$$\mathcal{B}_{A^*}(y, x) \leq \omega^*(\|y - x\|_x).$$

We are now ready to prove Proposition 5.5.1.

5.D.2 Proof of Proposition 5.5.1.

We start with Proposition 5.D.3, evaluated at $y = \mu^*$ and $x = \mu$:

$$\mathcal{B}_{A^*}(\mu^*, \mu) \leq \omega^*(\|\mu^* - \mu\|_\mu).$$

This hold if $\|\mu^* - \mu\|_\mu < 1$. Since ω^* is monotonically increasing, we can replace $\|\mu^* - \mu\|_\mu$ by its upper bound from Proposition 5.D.2,

$$\omega^*(\|\mu^* - \mu\|_\mu) \leq \omega^*\left(\frac{\|\mu^* - \mu\|_{\mu^*}}{1 - \|\mu^* - \mu\|_{\mu^*}}\right),$$

under the conditions that $\|\mu^* - \mu\|_{\mu^*} < 1$ (to satisfy the assumption of Proposition 5.D.2) and $\frac{\|\mu^* - \mu\|_{\mu^*}}{1 - \|\mu^* - \mu\|_{\mu^*}} < 1$ (to ensure that $\|\mu^* - \mu\|_\mu < 1$). Those two conditions holds if $\|\mu^* - \mu\|_{\mu^*} < 0.5$. Now, we use the bound (see figure 5.9)

$$\omega^*\left(\frac{t}{1-t}\right) \leq t^2, \quad 0 \leq t \leq 0.21,$$

and replace t by $\|\mu^* - \mu\|_{\mu^*}$. This finally gives the sequence of inequalities

$$\mathcal{B}_{A^*}(\mu^*, \mu) \leq \omega^*(\|\mu^* - \mu\|_{\mu}) \leq \omega^* \left(\frac{\|\mu^* - \mu\|_{\mu^*}}{1 - \|\mu^* - \mu\|_{\mu^*}} \right) \leq \|\mu^* - \mu\|_{\mu^*}^2,$$

that holds while $\|\mu^* - \mu\|_{\mu^*} < 0.21$, which is the desired result.

5.E Bias-Variance

In this section, we start from the notions of bias and variance introduced in Eq. (5.30). First, we prove that the bias of the MLE of a Gaussian variance decreases in $O(\frac{1}{n^2})$. Then we prove that assuming (5.31) holds, whether uniformly or in expectation, yields a convergence rate on the variance term.

5.E.1 Bias of a Gaussian Variance MLE

For a Gaussian variance model, the MLE follows a scaled $\chi^2(n)$ distribution. This means that

$$\frac{\mu^*}{\tilde{\mu}_n} = \frac{\tilde{\theta}_n}{\theta^*} = \mathbb{E}\left[\frac{\hat{\theta}_n}{\theta^*}\right] = \mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] = \mathbb{E}\left[\frac{n}{\chi^2(n)}\right] = \frac{n}{n-2}\,. \tag{5.115}$$

Consequently,

$$\mathcal{B}_{A^*}(\mu^*, \tilde{\mu}_n) = \frac{1}{2}\left(\frac{n}{n-2} - 1 + \log\frac{n-2}{n}\right) \tag{5.116}$$

$$= \frac{1}{n-2} + \frac{1}{2}\log\left(1 - \frac{2}{n}\right) \tag{5.117}$$

$$\leq \frac{1}{n-2} - \frac{1}{n} = \frac{2}{n(n-2)}, \tag{5.118}$$

so the bias of the MLE of a Gaussian Variance decreases like $O(\frac{1}{n^2})$.

5.E.2 Expectation of SMD's Variance Assumption

The first step is to notice the symmetrized Bregman can be expressed as an inner product between primal and dual parameters

$$\mathcal{S}_{A^*}(\mu, \bar{\mu}) = \langle \nabla A^*(\mu) - \nabla A^*(\bar{\mu}), \mu - \bar{\mu} \rangle = \langle \theta - \bar{\theta}, \mu - \bar{\mu} \rangle\,. \tag{5.119}$$

Now notice that (5.31) features $\mu := \hat{\mu}_{n+1} = \hat{\mu}_n - \gamma g(\hat{\theta}_n)$ stochastic and $\bar{\mu} := \bar{\mu}_{n+1} = \hat{\mu}_n - \gamma \nabla f(\hat{\theta}_n)$ deterministic such that $\mathbb{E}_g[\mu] = \bar{\mu}$. For such a pair of variables, the

expectation of the symmetrized Bregman corresponds to a covariance between primal and dual parameters

$$\mathbb{E}\left[\mathcal{S}_{A^*}(\mu, \mathbb{E}[\mu])\right] = \mathbb{E}\left[\langle \theta, \mu - \mathbb{E}[\mu]\rangle\right] = \underbrace{\mathbb{E}\left[\langle \theta, \mu\rangle\right] - \langle \mathbb{E}\left[\theta\right], \mathbb{E}\left[\mu\right]\rangle}_{\mathrm{Cov}(\theta,\mu)} = \mathbb{E}\left[\langle \theta - \mathbb{E}[\theta], \mu\rangle\right] = \mathbb{E}\left[\mathcal{S}_A(\theta, \mathbb{E}[\theta])\right] \tag{5.120}$$

The last equality holds by symmetry between the roles of A and A^*. Note that the middle covariance formulation is actually the one used by Hanzely and Richtárik (2021). Now, Eq. (5.30) defines the variance as

$$\mathbb{E}_{1:n}\left[\mathcal{B}_{A^*}(\tilde{\mu}_n, \hat{\mu}_n)\right] = \mathbb{E}_{1:n}\left[\mathcal{B}_A(\hat{\theta}_n, \mathbb{E}_{1:n}[\hat{\theta}_n])\right] \leq \mathbb{E}_{1:n}\left[\mathcal{S}_A(\hat{\theta}_n, \mathbb{E}_{1:n}[\hat{\theta}_n])\right] = \mathbb{E}_{1:n}\left[\mathcal{S}_{A^*}(\hat{\mu}_n, \mathbb{E}_{1:n}[\hat{\mu}_n])\right] \ , \tag{5.121}$$

where expectations $\mathbb{E}_{1:n}$ are on all samples $X_1, \ldots, X_n$, whereas (5.31) is written with

$$\mathbb{E}_n\left[\mathcal{S}_{A^*}(\hat{\mu}_n, \mathbb{E}_n[\hat{\mu}_n])\right] \leq \gamma^2 C \ , \tag{5.122}$$

where the expectation $\mathbb{E}_n$ is taken over only the last sample X_n, and the bound should hold uniformly over all $\hat{\mu}_{n-1}$. Taking the expectation over $\hat{\mu}_{n-1}$ instead gives

$$\mathbb{E}_{1:n}\left[\mathcal{S}_{A^*}(\hat{\mu}_n, \mathbb{E}_n[\hat{\mu}_n])\right] \leq \gamma^2 C \ . \tag{5.123}$$

The only difference with the right hand side of Eq. (5.121) is in the inner expectation. To overcome this difference, we need to plug in the form of $\hat{\mu}_n = \frac{n_0\mu_0 + \sum_i T_i}{n_0 + n}$. Notice that

$$\hat{\mu}_n - \mathbb{E}_n[\hat{\mu}_n] = \frac{T_n - \mu^*}{n_0 + n} \tag{5.124}$$

$$\implies \mathbb{E}_{1:n}\left[\mathcal{S}_{A^*}(\hat{\mu}_n, \mathbb{E}_n[\hat{\mu}_n])\right] = \frac{1}{n_0 + n}\mathbb{E}_{1:n}\left[\langle \hat{\theta}_n, T_1 - \mu^*\rangle\right] , \tag{5.125}$$

while

$$\hat{\mu}_n - \mathbb{E}_{1:n}[\hat{\mu}_n] = \frac{\sum_i (T_i - \mu^*)}{n_0 + n} \tag{5.126}$$

$$\implies \mathbb{E}_{1:n}\left[\mathcal{S}_{A^*}(\hat{\mu}_n, \mathbb{E}_{1:n}[\hat{\mu}_n])\right] = \frac{n}{n_0 + n}\mathbb{E}_{1:n}\left[\langle \hat{\theta}_n, T_1 - \mu^*\rangle\right] \ . \tag{5.127}$$

In the end, we get that the variance is dominated by n times the expectation of Eq. (5.31):

$$\mathbb{E}_{1:n}\left[\mathcal{B}_{A^*}(\tilde{\mu}_n, \hat{\mu}_n)\right] \leq n\mathbb{E}_{1:n}\left[\mathcal{S}_{A^*}(\hat{\mu}_n, \mathbb{E}_n[\hat{\mu}_n])\right] \ . \tag{5.128}$$

If assumption (5.31) holds, then we have

$$\mathbb{E}_{1:n}\left[\mathcal{B}_{A^*}(\tilde{\mu}_n, \hat{\mu}_n)\right] \leq n\gamma_n^2 C \in O(\frac{1}{n}) \ , \tag{5.129}$$

where we assumed $\gamma_n \in O(\frac{1}{n})$. In conclusion, assuming (5.31) holds uniformly or in expectation immediately implies a $O(\frac{1}{n})$ convergence rate on the variance.

5.F Review of SMD

We use this section to give more details on the (stochastic) mirror descent algorithm. We start with gradient descent with step-sizes γ. the update $\theta_{n+1} = \theta_n - \gamma \nabla f(\theta_n)$ can be viewed as the minimization of the linear approximation of f at θ_n $f(\theta) \approx f(\theta_n) + \langle \nabla f(\theta_n), \theta - \theta_n \rangle$, alongside with quadratic penalty scaled by $1/\gamma$:

$$\theta_{n+1} = \arg\min_{\theta} f(\theta_n) + \langle \nabla f(\theta_n), \theta - \theta_n \rangle + \frac{1}{\gamma}\frac{1}{2}\|\theta - \theta_n\|^2. \tag{5.130}$$

Mirror descent generalizes the above, using the Bregman divergence induced by a (Legendre) function A instead of the Euclidean norm as follow,

$$\theta_{n+1} = \arg\min_{\theta} f(\theta_n) + \langle \nabla f(\theta_n), \theta - \theta_n \rangle + \frac{1}{\gamma}\mathcal{B}_A(\theta, \theta_n). \tag{5.131}$$

Mirror descent coincides with gradient descent if $A(\theta) = \frac{1}{2}\|\theta\|^2$. As Eq. (5.131) is convex, the minimum is at a stationary point, found by taking the derivative and setting to 0, leading to the update θ_{n+1} satisfying

$$\nabla f(\theta_n) + \frac{1}{\gamma}(\nabla A(\theta_{n+1}) - \nabla A(\theta_n)) = 0 \quad \Longrightarrow \quad \nabla A(\theta_{n+1}) = \nabla A(\theta_n) - \gamma \nabla f(\theta_n). \tag{5.132}$$

Expressed with the dual parameters, we obtain $\mu_n = \nabla A(\theta_n)$, $\mu_{n+1} = \mu_n - \gamma \nabla f(\theta_n)$.

In our case, where the objective function is the (negative) log-likelihood of an exponential family, we have

$$f(\theta) = A(\theta) - \langle \frac{1}{n}\sum_{i=1}^{n} T(X_i), \theta \rangle. \tag{5.133}$$

Using Mirror descent with a step-size of 1 and the log-partition function A as the reference function gives

$$\mu_{n+1} = \mu_n - \nabla f(\theta_n) = \mu_n - (\nabla A(\theta_n) - \frac{1}{n}\sum_{i=1}^{n} T(x_i)) = \frac{1}{n}\sum_{i=1}^{n} T(x_i). \tag{5.134}$$

In a stochastic, online version where the linearization of the objective is obtained from iid samples, a decreasing step-size of $\gamma_n = 1/n$ recovers the "online" estimate of the MLE. The case of $\mu_1 = T(x_1)$ follows from the above, and in general, assuming it holds for μ_n,

$$\begin{aligned} \mu_{n+1} &= \mu_n - \gamma_n g(\theta_n) = \mu_n - \gamma_n(\mu_n - x_n) \\ &= (1 - \frac{1}{n})\mu_n + \frac{1}{n}T(x_n) = \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n-1} T(x_i) = \frac{1}{n}\sum_{i=1}^{n} T(x_i) + \frac{1}{n}T(x_n). \end{aligned} \tag{5.135}$$

The derivation in the main text gives the more general result, of using step-sizes of the form $1/(n + n_0)$ to recover online MAP estimation with a conjugate prior depending on n_0 and the initial estimate of the parameters θ_0.

6 Conclusion

The three contributions of this thesis deal with the interweaved topics of optimization and statistics. These three contributions can be summarized as

1. Variance reduction allows fast training of CRF, a particular class of conditional undirected graphical models that were previously hard to optimize. Thanks to duality, non-uniform sampling can be elegantly formalized and improved other strong methods.

2. For some simple classes of model, the causal model is faster to adapt to interventions than the anticausal one only when the intervention bears on the cause. However, our intuitions dictate that causal models should have some real-world advantages compared to non-causal ones. That may be why humans learn new rules so quickly. We may need more sophisticated models to instantiate this intuition in machine learning.

3. The maximum likelihood estimate of an exponential family can also be seen as the output of stochastic mirror descent. Furthermore, the KL divergence between the true and learned models is simply a Bregman divergence. Nevertheless, neither optimization nor statistics communities have found upper bounds on this quantity that apply to any sample size and families such a Gaussians. Finding such an upper bound may help non-Euclidean optimization reach new grounds.

This last contribution reveals that while exponential families are at the core of most machine learning techniques, some of their properties are yet to be understood. Throughout this thesis, we alternated between optimization and statistics perspectives, displaying the synergy between these two fields. Thanks to optimization tools, statistical models are becoming more powerful. Thanks to statistical models, we can probe into the abilities of optimization methods.

6.1 Future Work

Based on our last contribution, we identify two promising research directions.

1. finding high probability bounds for exponential family MAP thanks to the entropy being a self-concordant barrier, as proved by (Bubeck and Eldan, 2015). That would provide a general large sample result. A low sample result remains to be found.

2. Analyzing the convergence properties of stochastic mirror descent on self-concordant (barrier) losses. This might be possible thanks to the quadratic sandwich property of such functions, and it might be possible to find high probability convergence rate.

These research directions may help us understand fundamentals statistical models and design better optimization methods for barrier objectives.

Bibliography

A. Agarwal and H. Daumé. A geometric view of conjugate priors. *Machine learning*, 2010.

A. Aitken and H. Silverstone. On the estimation of statistical parameters. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 1942.

A. Anastasiou and G. Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 2017.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Y. F. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(1):310–342, 2017.

K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 2001.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 2010.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *NeurIPS*, 2013.

Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 2017.

O. Barndorff-Nielsen. *Information and Exponential Families*. Wiley, 1978. doi: https://doi.org/10.1002/9781118857281.ch8. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118857281.ch8.

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 2017.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 2003.

Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *ICLR*, 2020.

B. Birnbaum, N. R. Devanur, and L. Xiao. Distributed algorithms via gradient descent for fisher markets. In *ACM conference on Electronic commerce*, 2011.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

D. Braess and T. Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

S. Bubeck and R. Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. In *COLT*, 2015.

S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 2015.

A. Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

K. Chalupka, F. Eberhardt, and P. Perona. Estimating causal direction and confounding of two discrete variables. *arXiv preprint arXiv:1611.01504*, 2016.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In *NeurIPS*, 2001.

M. Collins, A. Globerson, T. Koo, X. Carreras, and P. L. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin Markov networks. *Journal of Machine Learning Research*, 2008.

G. F. Cooper and E. Herskovits. A bayesian method for constructing bayesian belief networks from databases. In *Uncertainty Proceedings 1991*. Elsevier, 1991.

D. Csiba, Z. Qu, and P. Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. In *ICML*, 2015.

S. Dasgupta and D. J. Hsu. On-line estimation with the multivariate Gaussian distribution. In *COLT*, 2007.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

L. P. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001. ISBN 0-387-95117-2.

R. D'Orazio, N. Loizou, I. Laradji, and I. Mitliagkas. Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize. *arXiv preprint arXiv:2110.15412*, 2021.

R.-A. Dragomir, M. Even, and H. Hendrikx. Fast stochastic Bregman gradient methods: Sharp analysis and variance reduction. In *ICML*, 2021.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

C. Dünner, T. Parnell, and M. Jaggi. Efficient use of limited-memory accelerators for linear learning on heterogeneous systems. In *NIPS*, 2017.

D. Eaton and K. Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pages 107–114, 2007.

D. Geiger, D. Heckerman, et al. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412–1440, 2002.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 2012.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In *ICML*, 2019.

F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *Computational Optimization and Applications*, 2021.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 2018a.

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 2018b.

Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*, 2012.

D. Janzing and B. Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10), 2010.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

S. Kakade, S. Shalev-Shwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf, 2009.

S. Kakade, O. Shamir, K. Sindharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *AISTATS*, 2010.

S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh. On learning distributions from their samples. In *COLT*, 2015.

N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, C. Pal, and Y. Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

D. Koller and N. Friedman. *Probabilistic Graphical Models*. The MIT Press, 2009.

Z. Kong and K. Chaudhuri. The expressive power of a class of normalizing flow models. In *AISTATS*. PMLR, 2020.

R. G. Krishnan, S. Lacoste-Julien, and D. Sontag. Barrier Frank-Wolfe for marginal inference. In *NIPS*, 2015.

F. Kunstner, R. Kumar, and M. Schmidt. Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent. In *AISTATS*, 2021.

S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien. Gradient-based neural DAG learning. In *ICLR*, 2020.

S. Lacoste-Julien, M. Schmidt, and F. R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method, 2012. Preprint. arXiv/1212.2002.

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, 2013.

J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

R. Le Priol, A. Piché, and S. Lacoste-Julien. Adaptive stochastic dual coordinate ascent for conditional random fields. In *UAI*, 2018.

R. Le Priol, R. Babanezhad, Y. Bengio, and S. Lacoste-Julien. An analysis of the adaptation speed of causal models. In *AISTATS*, 2021a.

R. Le Priol, F. Kunstner, D. Scieur, and S. Lacoste-Julien. Convergence rates for the map of an exponential family and stochastic mirror descent–an open problem. *arXiv preprint arXiv:2111.06826*, 2021b.

G. Lebanon and J. D. Lafferty. Boosting and maximum likelihood for exponential models. In *NIPS*, 2002.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *NIPS*, 2015.

N. Loizou, S. Vaswani, I. Hadj Laradji, and S. Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *AISTATS*, 2021.

H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 2018.

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *NeurIPS*, 2018.

T. L. Magnanti. Fenchel and lagrange duality are equivalent. 1974.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. CRC Press, 1989.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 2016.

C. N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, pages 65–80, 1982.

S. Muggleton. Alan turing and the development of artificial intelligence. *AI communications*, 27(1):3–10, 2014.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical science*, 2012.

A. S. Nemirosky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

A. Nemirovski, A. B. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 2009.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004a.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Applied Optimization. Springer US, 2004b.

Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Y. E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*. Applied Optimization. Springer, 2004c.

F. W. J. Olver, A. B. O. Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. V. S. B. R. Mille and, H. S. Cohl, and e. M. A. McClain. NIST digital library of mathematical functions. http://dlmf.nist.gov/, Release 1.1.3 of 2021-09-15, 2021. URL http://dlmf.nist.gov/.

A. Osokin, J.-B. Alayrac, I. Lukasewitz, P. Dokania, and S. Lacoste-Julien. Minding the gaps for block Frank-Wolfe optimization of structured SVMs. In *ICML*, 2016.

D. M. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 2021.

N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

J. Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. 1985.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. 1988.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl and E. Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 2014.

D. Perekrestenko, V. Cevher, and M. Jaggi. Faster coordinate descent via adaptive importance sampling. In *AISTATS*, 2017.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

D. Pfau. A generalized bias-variance decomposition for Bregman divergences. http://davidpfau.com/assets/generalized_bvd_proof.pdf, 2013. [Online; accessed February 23rd 2021].

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, Cambridge, 2nd edition, 1992.

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

F. Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

A. Rosenfeld, R. Zemel, and J. K. Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

D. Rothenhäusler, P. Bühlmann, and N. Meinshausen. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3):1688–1722, 2019.

N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, 2012.

M. Schmidt, R. Babanezhad, M. Ahmed, A. Defazio, A. Clifton, and A. Sarkar. Non-uniform stochastic average gradient method for training conditional random fields. In *AISTATS*, 2015.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *ICML*, 2012.

B. Schölkopf. Causality for machine learning. *arXiv:1911.10500*, 2019.

F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *NAACL*, 2003.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *arXiv:1309.2375*, 2013a.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14, 2013b.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 2016.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

C. Squires, Y. Wang, and C. Uhler. Permutation-based causal structure learning with unknown intervention targets. *arXiv preprint arXiv:1910.09007*, 2019.

S. M. Stigler. The epic story of maximum likelihood. *Statistical Science*, pages 598–620, 2007.

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 2019.

B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *NIPS*, 2004.

J. Tian and J. Pearl. Causal discovery from changes. In *UAI*, 2001.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

A. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

S. S. Varadhan. *Large deviations and applications*. SIAM, 1984.

T. Verma and J. Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305, 2008.

H. Wallach. Efficient training of conditional random fields. Master's thesis, University of Edinburgh, 2002.

S. v. d. Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 2011.

T. Yang, Q. Lin, and Z. Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *ICML*, 2013.

P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *ICML*, 2015.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *NeurIPS*, 2018.