CPSC 486/586: Probabilistic Machine Learning *Out: January 18, 2023*

Problem Set 1

Instructor: Andre Wibisono *Due: February 1, 2023*

(P1) Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the quadratic function: $f(x) = \frac{1}{2}\|x\|^2$.

(a) Write down the gradient descent algorithm with step size $\eta > 0$:

$$(\mathsf{GD}_f): \qquad x_{k+1} = x_k - \eta \nabla f(x_k).$$

Solve the recursion and determine how fast x_k converges to its limit x^* as $k \to \infty$. What is x^*? For which step size η does the conclusion hold?

(b) Write down the proximal gradient method with step size $\eta > 0$:

$$(\mathsf{PG}_f): \qquad x_{k+1} = \arg\min_{x\in\mathbb{R}^d} f(x) + \frac{1}{2\eta}\|x - x_k\|^2.$$

Solve the recursion and determine how fast x_k converges to its limit x^* as $k \to \infty$. For which step size η does the conclusion hold?

(c) Write down the solution to the gradient flow dynamics:

$$(\mathsf{GF}_f): \qquad \dot{X}_t = -\nabla f(X_t).$$

Write X_t in terms of X_0 and t, and determine how fast X_t converges to its limit x^* as $t \to \infty$. Suppose $t = \eta k$. How does your answer compare to part (a) and (b)?

(P2) Suppose we want to minimize $h\colon \mathbb{R} \to \mathbb{R}$ given by $h(x) = f(x) + g(x)$ where $f(x) = \frac{1}{2}(x-1)^2$ and $g(x) = \frac{1}{2}(x-1)^2$. Suppose we can only use the gradient descent $\mathsf{GD}_f, \mathsf{GD}_g$ or proximal gradient $\mathsf{PG}_f, \mathsf{PG}_g$ for f and g separately (but we cannot do GD_{f+g} or PG_{f+g}).

(a) For each algorithm below, write down its recursion explicitly, and determine its limit $x^*_\eta = \lim_{k\to\infty} x_k$.

i. $x_{k+1} = \mathsf{GD}_g \circ \mathsf{GD}_f(x_k)$

ii. $x_{k+1} = \mathsf{GD}_g \circ \mathsf{PG}_f(x_k)$

iii. $x_{k+1} = \mathsf{PG}_g \circ \mathsf{GD}_f(x_k)$

iv. $x_{k+1} = \mathsf{PG}_g \circ \mathsf{PG}_f(x_k)$

(b) For which combination above is the algorithm *consistent*, i.e. the limiting point x^*_η of the algorithm equal to the true minimizer $x^* = \arg\min_{x\in\mathbb{R}} f(x) + g(x)$? Can you explain why only certain combinations are consistent?

(P3) Let $X \sim \mathcal{N}(0, C)$ be a Gaussian random variable on $\mathbb{R}^d$ with mean $0 \in \mathbb{R}^d$ and covariance matrix $C \in \mathbb{R}^{d \times d}$. Assume $C \succ 0$ has eigenvalues $0 < \lambda_1 \le \cdots \le \lambda_d$. Evaluate the integral to compute the values below.

(a) Write down the following expression as a function of $\lambda_1, \dots, \lambda_d$:

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2} x^\top C^{-1} x} dx.$$

(b) For $\theta \in \mathbb{R}^d$, compute $\mathbb{E}[e^{\theta^\top X}]$. When is it finite?

(c) For $t > 0$, compute $\mathbb{E}[e^{t\|X\|^2}]$. When is it finite?

(d) Compute the (negative) entropy $H(\rho) = \mathbb{E}[\log \rho(X)]$. How does it scale with C?

(P4) Consider the noisy recursion:

$$x_{k+1} = (1 - \eta) x_k + \epsilon z_k$$

where $\eta > 0$ is step size and $\epsilon > 0$ is noise scale (usually $\eta, \epsilon \ll 1$), and $z_k \sim \mathcal{N}(0, I)$ an independent Gaussian random variable in $\mathbb{R}^d$. We start from any $x_0 \sim \rho_0$ to get $x_k \sim \rho_k$.

(a) Compute the mean $m_k = \mathbb{E}_{\rho_k}[x_k]$ and covariance matrix $C_k = \mathsf{Cov}_{\rho_k}(x_k)$ as a function of m_0, C_0, k. Determine how fast they converge to the limit $m_\infty = \lim_{k \to \infty} m_k$ and $C_\infty = \lim_{k \to \infty} C_k$.

(b) Determine what is the limiting distribution $\pi^* = \lim_{k \to \infty} \rho_k$ of the recursion. How does it depend on the step size and noise scale?

(c) Suppose $\epsilon = \eta$. What happens to the limiting distribution π^* for small $\eta \to 0$?

(d) Suppose $\epsilon = \sqrt{\eta}$. What happens to the limiting distribution π^* for small $\eta \to 0$?

(P5) (a) Recall the one-dimensional integration by parts formula (you may consult textbooks).

(b) Use the formula above to prove the following multi-dimensional integration by parts identity. Let $f \colon \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable function with $\lim_{\|x\| \to \infty} f(x) = 0$. Let $v \colon \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable vector field with $\lim_{\|x\| \to \infty} \|v(x)\| = 0$. Show that:

$$\int_{\mathbb{R}^d} \langle \nabla f(x), v(x) \rangle \, dx = - \int_{\mathbb{R}^d} f(x) \, \nabla \cdot v(x) \, dx \tag{1}$$

(c) Let $X \sim \mathcal{N}(0, I)$ be a Gaussian random variable in $\mathbb{R}^d$. Prove **Stein's identity**:

$$\mathbb{E}[\nabla f(X)] = \mathbb{E}[X \, f(X)]$$

If $X \sim \mathcal{N}(m, C)$ for some $m \in \mathbb{R}^d$, $C \succ 0$, how does the identity above change?

Additional questions for 586

(Q1) Let $0 < \epsilon \ll 1$ and y be a second-order perturbation of $x \in \mathbb{R}^d$ for some $u, v \in \mathbb{R}^d$:

$$y = x + \epsilon u + \epsilon^2 v$$

(a) Compute $\|y\|^2 - \|x\|^2$ as a polynomial in ϵ.

(b) Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be three-times differentiable. Compute $f(y) - f(x)$ up to $O(\epsilon^3)$ terms: $f(y) = f(x) + a(x)\epsilon + b(x)\epsilon^2 + O(\epsilon^3)$. Compute $a(x)$, $b(x)$ in terms of derivatives of $f(x)$.

(c) Let $\tilde{y} = x + \epsilon u$. Compute $f(\tilde{y}) - f(x)$ as a function of ϵ, and compare with $f(y) - f(x)$.

(d) Suppose we want to minimize f and we choose $u = -\nabla f(x)$, so up to first-order we are following gradient descent. What v should we choose to further decrease $f(y)$?

(Q2) Consider the recursion

$$x_{k+1} = x_k - \eta \nabla f(x_k) + b$$

where $0 < \eta \le \frac{1}{L}$ is step size and $b \in \mathbb{R}^d$. Assume f is α-strongly convex and L-smooth for some $\alpha > 0$, $L < \infty$.

(a) Show the map $F_\eta(x) = x - \eta \nabla f(x) + b$ is contractive: $\|F_\eta(x) - F_\eta(y)\| \le (1 - \alpha\eta)\|x - y\|$.

(b) Show that there is a unique limit point $x_\infty \in \mathbb{R}^d$ of F_η and that $x_k \to x_\infty$ exponentially fast: $\|x_k - x_\infty\| \le e^{-\alpha\eta k}\|x_0 - x_\infty\|$ for all $k \ge 0$. Compute x_∞ in terms of b, f, η.

(c) Let $x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$. Give an upper bound to $\|x_\infty - x^*\|$ in terms of b, α, η.

(Q3) Describe your research. What is the problem? How does it relate to probabilistic modeling or inference? (If you don't have research experience, you may describe a topic from a paper or a book.)