Statistical Mechanics of High Dimensional Inference
Supplementary Material

Madhu Advani
Surya Ganguli

Contents

1 Introduction

Here we provide both detailed derivations of results in the main paper as well as background information about classical results in statistics, to place our results on modern high dimensional inference in perspective.

2 Deriving inference error from the statistical physics of disordered systems

In this section, we derive consistency equations governing the mean squared error q_s between the true parameters $\mathbf{s^0} \in \mathbb{R}^P$ and our estimate of the parameters $\hat{\mathbf{s}} \in \mathbb{R}^P$. Einstein summation convention is assumed unless otherwise stated and measurements $(y_1, ..., y_N)$ are drawn as

$$y_\mu = X_{\mu j} s_j^0 + \epsilon_\mu. \tag{1}$$

Given knowledge of the measurements $\mathbf{y} \in \mathbb{R}^N$ and the measurement or input matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$ (the μ'th row of $\mathbf{X}$ is the measurement vector $\mathbf{x}^\mu$ in the main paper), but not the noise realizations ϵ_μ, which are drawn iid according to a probability density function P_ϵ, we consider M-estimators of the form:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \left[\sum_{\mu=1}^{N} \rho(y_\mu - X_{\mu j} s_j) + \sum_{j=1}^{P} \sigma(s_j) \right]. \tag{2}$$

The loss and regularizer ρ and σ will be assumed to be convex functions throughout this paper. In what follows we let $\mathbf{X} \in \mathbb{R}^{N \times P}$ be a random matrix with elements drawn iid from a normal distribution with variance $\frac{1}{P}$ as $X_{\mu j} \sim \mathcal{N}(0, \frac{1}{P})$, where the $\frac{1}{P}$ scaling serves to maintain an approximately unit norm for each measurement vector $\mathbf{x}^\mu$ so that both the noise $\boldsymbol{\epsilon}$ and signal $\mathbf{Xs^0}$ vectors have norms that are $O(N)$ if each component of the noise and signal are $O(1)$.

2.1 Replica equations at finite temperature

Result 1. ***Finite temperature RS relations*** *We let β be the inverse temperature and derive the RS (replica symmetric) equations in the high dimensional asymptotic (big-data) limit where $N, P \to \infty$ and the ratio of samples to variables has a finite limit $\frac{N}{P} \to \alpha$. We find that average squared residual error of the energy minimization problem (4), reduces to a set of 4 equations relating scalar order parameters $q_d, q_s, \Lambda_\rho, \Lambda_\sigma$*

$$\frac{q_d}{\Lambda_\sigma^2} = \frac{\alpha}{\Lambda_\rho^2} \left(\left\langle\!\!\left\langle \left\langle d - \epsilon_{q_s} \right\rangle_{P_\rho}^2 \right\rangle\!\!\right\rangle_{\epsilon, z_\epsilon} \right) \qquad\Bigg|\qquad q_s = \left\langle\!\!\left\langle \left\langle s - s_0 \right\rangle_{P_\sigma}^2 \right\rangle\!\!\right\rangle_{s^0, z_s}$$

$$\frac{1}{\Lambda_\sigma} = \frac{\alpha}{\Lambda_\rho^2} \left(\Lambda_\rho - \left\langle\!\!\left\langle \left\langle (\delta d)^2 \right\rangle_{P_\rho} \right\rangle\!\!\right\rangle_{\epsilon, z_\epsilon} \right) \qquad\Bigg|\qquad \Lambda_\rho = \left\langle\!\!\left\langle \left\langle (\delta s)^2 \right\rangle_{P_\sigma} \right\rangle\!\!\right\rangle_{s^0, z_s}$$

$$P_\rho(d|\epsilon_{q_s}) \propto e^{-\frac{(d-\epsilon_{q_s})^2}{2\Lambda_\rho} - \beta\rho(d)}, \qquad P_\sigma(s|s_{q_d}^0) \propto e^{-\frac{1}{2\Lambda_\sigma}\left(s - s_{q_d}^0\right)^2 - \beta\sigma(s)} \tag{3}$$

z_s, z_ϵ are independent random unit gaussians, and s^0, ϵ are random variables drawn from the parameter and noise probability densities P_s and P_ϵ respectively, and we define the effective noise $\epsilon_{q_s} = \epsilon + \sqrt{q_s} z_\epsilon$ and effective parameters $s_{q_d}^0 = s^0 + \sqrt{q_d} z_s$. We define thermal fluctuations from the mean $\delta s = s - \langle s \rangle$ and $\delta d = d - \langle d \rangle$.

Derivation of Result 1. We now introduce a physical system which reduces to the above inference problem in the zero temperature limit. The energy function of such a system is

$$E(\mathbf{u}) = \sum_{\mu=1}^{N} \rho\left(\epsilon_\mu - X_{\mu j} u_j\right) + \sum_{j=1}^{P} \sigma(s_j^0 - u_j), \tag{4}$$

where we have made a change of variable to write the problem in terms of the residual error $\mathbf{u}$ defined as $\mathbf{u} = \mathbf{s^0} - \mathbf{s}$. The Gibbs distribution on the residual error $\mathbf{u}$ is $P_{\mathrm{G}}(\mathbf{u}) = \frac{e^{-\beta E(\mathbf{u})}}{Z}$. We then compute the average Free Energy of the system by applying the replica trick based on the identity $\lim_{n \to 0} \frac{\langle\langle Z^n \rangle\rangle - 1}{n} = \langle\langle \ln(Z) \rangle\rangle$. Note that $\mathbf{X}, \mathbf{s^0}, \boldsymbol{\epsilon}$ here play the role of quenched disorder. To proceed with the replica method, we first compute:

$$\langle\langle Z^n \rangle\rangle_{\epsilon, X, s^0} = \left\langle\!\!\left\langle \int \prod_{a=1}^{n} \mathbf{du^a} e^{-\beta \sum_{a=1}^{n} \left(\sum_{\mu=1}^{N} \rho(X_{\mu k} u_k^a + \epsilon_\mu) + \sum_{j=1}^{P} \sigma(s_j^0 - u_j^a) \right)} \right\rangle\!\!\right\rangle_{\epsilon, X, s^0}. \tag{5}$$

Note that, since $X_{\mu j} \sim \mathcal{N}(0, \frac{1}{P})$, for fixed $\mathbf{u^a}$, $b_\mu^a = \sum_{j=1}^{P} X_{\mu j} u_j^a$, is a gaussian with a covariance structure $\left\langle\!\left\langle b_\mu^a b_\nu^b \right\rangle\!\right\rangle_X = Q_{ab} \delta_{\mu\nu}$ so that

$$\langle\langle\, Z^n \,\rangle\rangle_{\epsilon,X,s^0} = \int \mathbf{ds^0} P_s(\mathbf{s^0}) \int \prod_{ab} dQ_{ab} \int \prod_{a} \mathbf{du^a} \prod_{ab} \delta(u^a \cdot u^b - PQ_{ab}) e^{-\beta \sum_{j=1}^{P} \sum_a \sigma(s_j^0 - u_j^a)}$$
$$\int \prod_{\mu=1}^{N} d\epsilon_\mu P_\epsilon(\epsilon_\mu) \int \frac{1}{|Q|^N} \prod_{\mu a} \frac{db_\mu^a}{\sqrt{2\pi}} e^{\sum_{\mu=1}^{N} -\beta \sum_a \rho(b_\mu^a + \epsilon_\mu) - \frac{1}{2} \sum_{ab} b_\mu^a Q_{ab}^{-1} b_\mu^b}, \tag{6}$$

where $|Q|$ denotes the determinant of Q. By noting that part of the integral factorizes in μ we have

$$\langle\langle\, Z^n \,\rangle\rangle_{\epsilon,X,s^0} = \int \prod_{ab} dQ_{ab} e^{-NE(Q)} \int \mathbf{ds^0} \prod_{a} \mathbf{du^a} P_s(\mathbf{s^0}) e^{-\beta \sum_{j=1}^{P} \sum_a \sigma(s_j^0 - u_j^a)} \prod_{ab} \delta\left(\mathbf{u}^a \cdot \mathbf{u}^a - PQ_{ab}\right), \tag{7}$$

where

$$E(Q) = -\ln\left(\int d\epsilon P_\epsilon(\epsilon) \int \prod_a db^a \frac{1}{|Q|} e^{-\frac{1}{2} b^a Q_{ab}^{-1} b^b} e^{-\beta \sum_{a=1}^{n} \rho(\epsilon + b^a)}\right). \tag{8}$$

We then replace our delta functions with integrals over dummy variables $\hat{Q}$ on the imaginary axis

$$\langle\langle\, Z^n \,\rangle\rangle_{\epsilon,X,s^0} = \int \prod_{ab} dQ_{ab} d\hat{Q}_{ab} e^{-NE(Q) - \sum_{ab} P\hat{Q}_{ab} Q_{ab}} \int \mathbf{ds^0} \prod_a \mathbf{du^a} P_s(\mathbf{s^0}) e^{-\beta \sum_{j=1}^{P} \sum_a \sigma(s_j^0 - u_j^a) + \sum_{ab} \hat{Q}_{ab} \mathbf{u}^a \cdot \mathbf{u}^b}. \tag{9}$$

We can simplify this by noting the factorization of the RHS to find

$$\langle\langle\, Z^n \,\rangle\rangle_{\epsilon,X,s^0} = \int \prod_{ab}^{n} dQ_{ab} d\hat{Q}_{ab} e^{-N\left(E(Q) + \frac{1}{\alpha} \sum_{ab} \hat{Q}_{ab} Q_{ab} - \frac{1}{\alpha} S(\hat{Q})\right)}, \tag{10}$$

where

$$S(\hat{Q}) = \ln\left(\int ds^0 \prod_a^n du^a P_s(s^0) e^{-\beta \sum_a \sigma(s^0 - u^a) + \sum_{ab} \hat{Q}_{ab} u^a u^b}\right). \tag{11}$$

We now apply the replica symmetric (RS) assumption about the form of $\mathbf{Q}$, and a corresponding form on $\hat{\mathbf{Q}}$. For $a \neq b$, $Q_{ab} = Q_0$, $Q_{aa} = Q_1$, $Q_{ab} = \hat{Q}_0$, and $Q_{aa} = \hat{Q}_1$. It is also useful to define the differences between the replicas $\Delta = Q_1 - Q_0$ and $\hat{\Delta} = \hat{Q}_1 - \hat{Q}_0$. The RS assumption implies that

$$b^a = \sqrt{\Delta} x^a + \sqrt{Q_0} z, \tag{12}$$

where x^a, z are unit gaussian random variables and $\Delta = Q_1 - Q_0$, so that

$$e^{-E(Q)} = \left\langle\!\!\left\langle \int \textstyle\prod_{a=1}^{n} \mathcal{D}x^a e^{-\beta \sum_{a=1}^{n} \rho\left(\epsilon + \sqrt{Q_0} z + \sqrt{\Delta} x^a\right)} \right\rangle\!\!\right\rangle_{z,\epsilon}. \tag{13}$$

Here $\mathcal{D}x = dx \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ is a gaussian integral. Noting the factorization over n replicas, this can be simplified as

$$e^{-E(Q)} = \left\langle\!\left\langle\, e^{-n\Lambda(z,\epsilon)} \,\right\rangle\!\right\rangle_{z,\epsilon}, \tag{14}$$

where

$$\Lambda(z,\epsilon) = -\ln\left(\int \mathcal{D}x e^{-\beta\rho\left(\epsilon + \sqrt{Q_0} z + \sqrt{\Delta} x\right)}\right). \tag{15}$$

Similarly we can factorize

$$e^{S(\hat{Q})} = \int ds^0 \prod_a^n du^a P_s(s^0) e^{-\beta \sum_a \sigma(s^0 - u^a) + \hat{Q}_0 (\sum_a u^a)^2 + \hat{\Delta} \sum_a (u^a)^2}, \tag{16}$$

where $\hat{\Delta} = \hat{Q}_1 - \hat{Q}_0$. To factorize the above expression, we exploit the identity $\langle\langle\, e^{\gamma\eta} \,\rangle\rangle_\eta = e^{\frac{\gamma^2}{2}}$ for unit gaussian η. Letting $\gamma = \sqrt{2\hat{Q}_0} \sum_a u^a$

$$e^{S(\hat{Q})} = \left\langle\!\!\left\langle \int ds^0 P_s(s^0) \textstyle\prod_a du^a e^{\sum_a \left[\eta \sqrt{2\hat{Q}_0} u^a + \hat{\Delta} (u^a)^2 - \beta\sigma(s^0 - u^a)\right]} \right\rangle\!\!\right\rangle_\eta. \tag{17}$$

It follows that,

$$e^{S(\hat{Q})} = \left\langle\!\!\left\langle\, e^{n\chi(\eta, s^0)} \,\right\rangle\!\!\right\rangle_{\eta, s^0}, \tag{18}$$

$$\chi(\eta, s^0) = \ln\left(\int du e^{\eta\sqrt{2\hat{Q}_0}u + \hat{\Delta}u^2 - \beta\sigma(s^0 - u)}\right). \tag{19}$$

To determine F, we calculate the low n Taylor expansion of $\langle\langle Z^n \rangle\rangle$, to do this we compute:

$$\lim_{n\to 0}\frac{\partial E}{\partial n} = \langle\langle \Lambda(z,\epsilon) \rangle\rangle_{z,\epsilon}, \tag{20}$$

$$\lim_{n\to 0}\frac{\partial S}{\partial n} = \left\langle\left\langle \chi(\eta, s^0) \right\rangle\right\rangle_{\eta,s^0}, \tag{21}$$

$$\lim_{n\to 0}\frac{\partial\left(Q_{ab}\hat{Q}_{ab}\right)}{\partial n} = Q_1\hat{Q}_1 - Q_0\hat{Q}_0 = Q_0\hat{\Delta} + \hat{Q}_0\Delta + \hat{\Delta}\Delta. \tag{22}$$

Using these limits to compute the Free energy density given by the saddle point in (10)

$$\tilde{F} = \frac{-\beta F}{N} = \min_{Q_0,\hat{Q}_0,\Delta,\hat{\Delta}}\left[\langle\langle \Lambda(z,\epsilon) \rangle\rangle_{z,\epsilon} + \frac{1}{\alpha}\left(Q_0\hat{\Delta} + \hat{Q}_0\Delta + \hat{\Delta}\Delta\right) - \frac{1}{\alpha}\left\langle\left\langle \chi(\eta, s^0) \right\rangle\right\rangle_{\eta,s^0}\right]. \tag{23}$$

Minimizing with respect to each of the order parameters $Q_0, \hat{Q}_0, \Delta, \hat{\Delta}$, we find the following set of coupled equations defining the order parameters:

$$\hat{\Delta} = -\alpha\frac{\partial\langle\langle \Lambda \rangle\rangle_{z,\epsilon}}{\partial Q_0} \tag{24}$$

$$\hat{Q}_0 + \hat{\Delta} = -\alpha\frac{\partial\langle\langle \Lambda \rangle\rangle_{z,\epsilon}}{\partial \Delta} \tag{25}$$

$$\Delta = \frac{\partial\langle\langle \chi \rangle\rangle_{\eta,s^0}}{\partial\hat{Q}_0} \tag{26}$$

$$Q_0 + \Delta = \frac{\partial\langle\langle \chi \rangle\rangle_{\eta,s^0}}{\partial\hat{\Delta}}. \tag{27}$$

By making a change of variables, we can write

$$\Lambda(z,\epsilon) = -\ln\left(\frac{1}{\sqrt{\Delta}}\int dy e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2 - \beta\rho(y)}\right), \tag{28}$$

so that

$$\frac{\partial\langle\langle \Lambda \rangle\rangle_{z,\epsilon}}{\partial Q_0} = -\frac{1}{2\Delta\sqrt{Q_0}}\left\langle\left\langle z\frac{\int dy(y-\sqrt{Q_0}z-\epsilon)e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}{\int dy e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}\right\rangle\right\rangle_{z,\epsilon}. \tag{29}$$

Using the fact that for a gaussian random variable z, $\langle\langle zf(z) \rangle\rangle_z = \langle\langle f'(z) \rangle\rangle_z$, we can re-write this as

$$\frac{\partial\langle\langle \Lambda \rangle\rangle_{z,\epsilon}}{\partial Q_0} = -\frac{1}{2\Delta^2}\left(\left\langle\left\langle \frac{\int dy(y-\sqrt{Q_0}z-\epsilon)^2 e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}{\int dy e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}\right\rangle\right\rangle_{z,\epsilon} - \left\langle\left\langle\left(\frac{\int dy(y-\sqrt{Q_0}z-\epsilon)e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}{\int dy e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}\right)^2\right\rangle\right\rangle_{z,\epsilon} - \Delta\right). \tag{30}$$

Defining $\epsilon_{Q_0} = \epsilon + \sqrt{Q_0}z$ yields

$$\hat{\Delta} = \frac{\alpha}{2\Delta^2}\left\langle\left\langle \left\langle (y-\epsilon_{Q_0})^2 \right\rangle_{P_\rho(y)} - \left\langle (y-\epsilon_{Q_0}) \right\rangle_{P_\rho(y)}^2 \right\rangle\right\rangle_{\epsilon_{Q_0}} - \frac{\alpha}{2\Delta}, \tag{31}$$

where the $\langle\,.\,\rangle_{P_\rho(y)}$ denotes average over a Gibbs distributions $P_\rho(y) \propto e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}$. Similarly, we find

$$\frac{\partial\langle\langle \Lambda \rangle\rangle_{z,\epsilon}}{\partial \Delta} = -\frac{1}{\Delta^2}\left\langle\left\langle \frac{\int dy(y-\sqrt{Q_0}z-\epsilon)^2 e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}{\int dy e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}}\right\rangle\right\rangle_{z,\epsilon} + \frac{1}{2\Delta}, \tag{32}$$

implying

$$\hat{Q}_0 + \hat{\Delta} = \frac{\alpha}{2\Delta^2}\left\langle\left\langle \left\langle (y-\epsilon_{Q_0})^2 \right\rangle_{P_\rho(y)} \right\rangle\right\rangle_{\epsilon_{Q_0}} - \frac{\alpha}{2\Delta}. \tag{33}$$

We can split this expectation into a variance term and bias term to yield

$$\hat{Q}_0 = \frac{\alpha}{2\Delta^2}\left\langle\left\langle \left\langle y-\epsilon_{Q_0} \right\rangle_{P_\rho(y)}^2 \right\rangle\right\rangle_{\epsilon_{Q_0}}. \tag{34}$$

We can also compute order parameters (Δ, Q_0), and find

$$\Delta = \left\langle\left\langle \eta\frac{\int du\frac{u}{\sqrt{2\hat{Q}_0}}e^{\eta\sqrt{2\hat{Q}_0}u+\hat{\Delta}u^2-\beta\sigma(s^0-u)}}{\int du e^{\eta\sqrt{2\hat{Q}_0}u+\hat{\Delta}u^2-\beta\sigma(s^0-u)}}\right\rangle\right\rangle_{\eta,s^0}, \tag{35}$$

where by using the property of gaussian integrals to take the derivative with respect to η, we find

$$\Delta = \left\langle\!\!\left\langle \langle u^2 \rangle_{P_\sigma(u)} - \langle u \rangle^2_{P_\sigma(u)} \right\rangle\!\!\right\rangle_{\eta,s^0}, \tag{36}$$

where $P_\sigma(u) \propto e^{\hat{\Delta}\left(u^2+2\sqrt{\frac{\hat{Q}_0}{2\hat{\Delta}^2}}u\eta\right)-\beta\sigma(s^0-u)} \propto e^{\hat{\Delta}\left(u+\sqrt{\frac{\hat{Q}_0}{2\hat{\Delta}^2}}\eta\right)^2-\beta\sigma(s^0-u)}$. Similarly, minimizing with respect to $\hat{\Delta}$, we find

$$Q_0 + \Delta = \left\langle\!\!\left\langle \frac{\int du u^2 e^{\eta\sqrt{2\hat{Q}_0}u+\hat{\Delta}u^2-\beta\sigma(s^0-u)}}{\int du e^{\eta\sqrt{2\hat{Q}_0}u+\hat{\Delta}u^2-\beta\sigma(s^0-u)}} \right\rangle\!\!\right\rangle_{\eta,s^0} = \left\langle\!\!\left\langle \langle u^2 \rangle_{P_\sigma(u)} \right\rangle\!\!\right\rangle_{\eta,s^0}, \tag{37}$$

splitting this into a bias and variance component yields

$$Q_0 = \left\langle\!\!\left\langle \langle u \rangle^2_{P_\sigma(u)} \right\rangle\!\!\right\rangle_{\eta,s^0}. \tag{38}$$

These relations, which hold for arbitrary temperature, may be written as

$$\hat{Q}_0 = \frac{\alpha}{2\Delta^2}\left(\left\langle\!\!\left\langle \langle y - \epsilon_{Q_0} \rangle^2_{P_\rho(y)} \right\rangle\!\!\right\rangle_{\epsilon_{Q_0}}\right), \tag{39}$$

$$\hat{\Delta} = \frac{\alpha}{2\Delta^2}\left(\left\langle\!\!\left\langle \langle \delta y^2 \rangle_{P_\rho(y)} \right\rangle\!\!\right\rangle_{\epsilon_{Q_0}} - \Delta\right), \tag{40}$$

$$Q_0 = \left\langle\!\!\left\langle \langle u \rangle^2_{P_\sigma(u)} \right\rangle\!\!\right\rangle_{s^0,\eta}, \tag{41}$$

$$\Delta = \left\langle\!\!\left\langle \langle \delta s^2 \rangle_{P_\sigma(u)} \right\rangle\!\!\right\rangle_{s^0,\eta}, \tag{42}$$

where $q_d = \frac{\hat{Q}_0}{2\hat{\Delta}^2}$, $\delta u = u - \langle u \rangle$, $\delta y = y - \langle y \rangle$, and

$$P_\rho(y) \propto e^{-\frac{1}{2\Delta}\left(y-\sqrt{Q_0}z-\epsilon\right)^2-\beta\rho(y)}, \qquad P_\sigma(u) \propto e^{\hat{\Delta}\left(s-s^0+\sqrt{q_d}\eta\right)^2-\beta\sigma(s)}. \tag{43}$$

We make the simplifying change of variables $\Delta \to \Lambda_\rho$, $\hat{\Delta} \to \frac{-1}{2\Lambda_\sigma}$, $Q_0 \to q_s$, and $\hat{Q}_0 \to \frac{q_d}{2\Lambda_\sigma^2}$. We also relabel our gaussian unit noise $z \to z_\epsilon$, $\eta \to z_s$ to match the main paper.

2.2 Replica equations in the low temperature limit

Result 2. ***The low temperature limit:***
In the low temperature limit $\beta \to \infty$, the energy minimization problem and M-estimation problem are equivalent. We seek solutions to the replica equations where the order parameters scale with β as $\Lambda_\rho = \frac{\lambda_\rho}{\beta}$, $\Lambda_\sigma = \frac{\lambda_\sigma}{\beta}$, so that $\lambda_\sigma, \lambda_\rho, q_s, q_d$ are O(1) scalars and we find that the probability distributions for d and s simplify in the low temperature limit to:

$$P_\rho(d|\epsilon_{q_s}) = \mathcal{N}\left(\hat{\epsilon}, \frac{\lambda_\rho \hat{\epsilon}'(\epsilon_{q_s})}{\beta}\right), \tag{44}$$

$$\hat{\epsilon}(\epsilon_{q_s}) = \mathcal{P}_{\lambda_\rho}[\,\rho\,](\epsilon_{q_s}), \tag{45}$$

$$P_\sigma(s|s^0_{q_d}) = \mathcal{N}\left(\hat{s}, \frac{\lambda_\sigma \hat{s}'(s^0_{q_d})}{\beta}\right), \tag{46}$$

$$\hat{s}(s^0_{q_d}) = \mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0_{q_d}). \tag{47}$$

where $\mathcal{P}_\lambda[\,f\,]$ is the proximal map function (see Appendix C for more information) defined as

$$\mathcal{P}_\lambda[\,f\,](x) = \arg\min_y \left[\frac{(x-y)^2}{2\lambda} + f(y)\right], \tag{48}$$

and $\epsilon_{q_s} = \epsilon + \sqrt{q_s}z_\epsilon, s^0_{q_d} = s^0 + \sqrt{q_d}z_s$ are random variables corresponding to the corrupted versions of the noise and signal where z_ϵ, z_s defined as zero mean, unit variance gaussian noise.

Derivation—.

$$P_\rho(d|\epsilon_{q_s}) \propto e^{-\beta\left[\frac{(d-\epsilon_{q_s})^2}{2\lambda_\rho}+\rho(d)\right]} \tag{49}$$

$$P_\sigma(s|s^0_{q_d}) \propto e^{-\beta\left[\frac{(s-s^0_{q_d})^2}{2\lambda_\sigma}+\sigma(s)\right]} \tag{50}$$

If we define $\hat{\epsilon}$ and $\hat{s}$ to be the mode (arg max) of distributions P_ρ and P_σ respectively, which is equivalent to maximizing the argument of the exponentials, then

$$\hat{\epsilon} = \mathcal{P}_{\lambda_\rho}[\,\rho\,](\epsilon_{q_s}), \tag{51}$$

$$\hat{s} = \mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0_{q_d}). \tag{52}$$

Taylor expanding P_ρ and P_σ about these maxima yields

$$P_\rho(d|\epsilon_{q_s}) \propto e^{-\frac{\beta}{2}\left[\frac{1}{\lambda_\rho}+\rho''(\hat{\epsilon})\right](d-\hat{\epsilon})^2}, \tag{53}$$

$$P_\sigma(s|s^0_{q_d}) \propto e^{-\frac{\beta}{2}\left[\frac{1}{\lambda_\sigma}+\sigma''(\hat{s})\right](s-\hat{s})^2}. \tag{54}$$

We can simplify the terms containing second derivatives in (53) and (54) as follows. Note that by the definition of the proximal map, it satisfies an interesting property: by differentiating the argument in the RHS of (48) and evaluating it at the proximal point (LHS of (48)), we necessarily get zero. Applying this property to (51) yields:

$$\frac{\hat{\epsilon}-\epsilon_{q_s}}{\lambda_\rho}+\rho'(\hat{\epsilon})=0. \tag{55}$$

Differentiating with respect to $\hat{\epsilon}$, we find

$$\frac{1}{\lambda_\rho}+\rho''(\hat{\epsilon})=\frac{1}{\lambda_\rho \hat{\epsilon}'(\epsilon_{q_s})}. \tag{56}$$

Substituting this form into (53) and similarly for (54), it follows that

$$P_\rho(d|\epsilon_{q_s}) \propto e^{-\frac{\beta(d-\hat{\epsilon})^2}{2\lambda_\rho \hat{\epsilon}'(\epsilon_{q_s})}}, \tag{57}$$

$$P_\sigma(s|s^0_{q_d}) \propto e^{-\frac{\beta(s-\hat{s})^2}{2\lambda_\sigma \hat{s}'(s^0_{q_d})}}. \tag{58}$$

Result 3. ***M-estimator error predictions*** *In the regularized case, consider any convex M-estimator defined by a loss function, regularizer pair (ρ,σ) where the parameters $(s^0_1,...,s^0_P)$ and noise $(\epsilon_1,...,\epsilon_N)$ are iid random variables. The associated zero temperature replica symmetric equations simplify to:*

$$\frac{q_d}{\lambda_\sigma^2}=\frac{\alpha}{\lambda_\rho^2}\Big\langle\Big\langle\,(\hat{\epsilon}-\epsilon_{q_s})^2\,\Big\rangle\Big\rangle, \tag{59}$$

$$q_s=\Big\langle\Big\langle\,(\hat{s}-s^0)^2\,\Big\rangle\Big\rangle_{s^0,z_s}, \tag{61}$$

$$\frac{1}{\lambda_\sigma}=\frac{\alpha}{\lambda_\rho}\langle\langle\,(1-\hat{\epsilon}'(\epsilon_{q_s}))\,\rangle\rangle, \tag{60}$$

$$\lambda_\rho=\lambda_\sigma\Big\langle\Big\langle\,\hat{s}'(s^0_{q_d})\,\Big\rangle\Big\rangle_{s^0,z_s}, \tag{62}$$

where

$$\hat{\epsilon}(\epsilon_{q_s})=\mathcal{P}_{\lambda_\rho}[\,\rho\,](\epsilon_{q_s}), \tag{63}$$

$$\hat{s}(s^0_{q_d})=\mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0_{q_d}). \tag{64}$$

Derivation.— This result follows from the substituting the low temperature form of P_ρ and P_σ in Result 2 into the finite temperature consistency equations in Result 1 and taking the zero temperature limit.

We note that these replica equations have a well-defined relationship to the original M-estimation problem in (2). Consider for example pairs of estimated and true signal components: $(\hat{s}_i, s^0_i)$ for $i=1,\dots P$ in the original high-dimensional inference problem. We can obtain an empirical histogram of such joint pairs, that converges as $P\to\infty$ to a joint distribution $P(\hat{s}_i, s^0_i)$. This distribution is self-averaging - i.e. does not depend on the detailed realization of signal, measurements, or noise. The replica theory makes a mean-field prediction for this self-averaging joint distribution through (64), as described in the main paper. This joint distribution arises as an auxiliary scalar inference problem in which gaussian noise of variance q_d is added to s^0, and this noise is cleaned up by a proximal descent step to obtain an estimate $\hat{s}$. Integrating out the noise yields the mean field prediction $P_{MF}(\hat{s}_i, s^0_i)$.

A similar statement holds from the perspective of noise as opposed to signal estimation. Any inference procedure that infers an estimated signal $\hat{\mathbf{s}}$ also implicitly estimates the noise through $\hat{\boldsymbol{\epsilon}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{s}}$, thus yielding a decomposition of the measurement outcome $\mathbf{y}$ into a sum of estimated signal contribution $\boldsymbol{X}\hat{\boldsymbol{s}}$ and estimated noise contribution $\hat{\boldsymbol{\epsilon}}$. Just as we did for the estimated signal and the true signal, we could also consider the joint distribution of components of the estimated noise and true noise, $(\hat{\epsilon}_\mu, \epsilon_\mu)$ for $\mu=1,\dots,N$. Again in the $N\to\infty$ limit the empirical joint histogram of these quantities converges to a self-averaging joint distribution $P(\hat{\epsilon}_\mu,\epsilon_\mu)$. The replica theory makes a mean field prediction for this joint distribution through (63), also described in the main paper. This joint distribution arises as an auxiliary scalar inference problem in which gaussian noise of variance q_s is added to the true noise ϵ, and this additional noise is cleaned up by a proximal descent step to obtain an estimated noise $\hat{\epsilon}$. Integrating out the Gaussian noise yields the mean field prediction $P_{MF}(\hat{\epsilon}_\mu,\epsilon_\mu)$. These distributional predictions can be derived by following a standard replica method in which the above steps are modified by the introduction of a delta function picking off one component of either the estimated signal or noise. See [1] section 8.2 for an example of this modification.

Result 4. *The special case of unregularized M-estimation, for which $\sigma=0$, follows directly from Result 3. In this case:*

$$\alpha\Big\langle\Big\langle\,(\hat{\epsilon}-\epsilon_{q_s})^2\,\Big\rangle\Big\rangle=q_s \tag{65}$$

$$\alpha\langle\langle\,1-\hat{\epsilon}'(\epsilon_{q_s})\,\rangle\rangle=1 \tag{66}$$

where

$$\hat{\epsilon}(\epsilon_{q_s})=\mathcal{P}_{\lambda_\rho}[\,\rho\,](\epsilon_{q_s}). \tag{67}$$

Using the relation between the Moreau envelope and proximal map, we can also write these equations in the form:

$$\alpha\Big\langle\Big\langle\,\big(\lambda_\rho\mathcal{M}_{\lambda_\rho}[\,\rho\,]'(\epsilon_{q_s})\big)^2\,\Big\rangle\Big\rangle=q_s \tag{68}$$

$$\alpha\Big\langle\Big\langle\,\lambda_\rho\mathcal{M}_{\lambda_\rho}[\,\rho\,]''(\epsilon_{q_s})\,\Big\rangle\Big\rangle=1 \tag{69}$$

the mean square M-estimator error q_s as well as the order parameter λ_ρ are may be computed as the solution to this pair of equations, this result is in agreement with the findings of previous work [2].

Derivation of Result 4.

The choice $\sigma = 0$ yields a linear map $\mathcal{P}_{\lambda_\sigma}[\,\sigma\,](x) = x$. Substituting this into (62) yields $\lambda_\rho = \lambda_\sigma$, and into (61) yields $q_s = q_d$. Using this simplification equations (59) and (60) become (65) and (66). Then, direct substitution of the relation between the proximal map and the Moreau envelope (74) derived in Appendix C.1 yields (68) and (69). Moreover, dividing (68) by the square of (69) yields the form of the replica equations exhibited in the main paper.

2.3 Moreau envelope formulation of the replica equations

While the above replica symmetric equations (59-61) provide a nice characterization of the mean field distributions $P_{\mathrm{MF}}(\hat{\epsilon},\epsilon)$ and $P_{\mathrm{MF}}(\hat{s},s^0)$, it will also be useful to formulate these replica symmetric equations in terms of Moreau envelopes, instead of proximal maps. This reformulation will simplify derivations of bounds on inference error in section 4.2 and derivations of optimal inference procedures in section 5.2. To achieve this reformulation, we systematically apply the relation between proximal maps and Moreau envelopes (74) derived in Appendix C.1. This yields the result:

Result 5. ***M-estimator Formulation***

$$\frac{q_d}{\lambda_\sigma^2} = \alpha\Big\langle\Big\langle \left(\mathcal{M}_{\lambda_\rho}[\,\rho\,]'(\epsilon_{q_s})\right)^2 \Big\rangle\Big\rangle \tag{70}$$

$$\frac{1}{\lambda_\sigma} = \alpha\Big\langle\Big\langle \mathcal{M}_{\lambda_\rho}[\,\rho\,]''(\epsilon_{q_s}) \Big\rangle\Big\rangle \tag{71}$$

$$q_s + q_d\left(1 - 2\frac{\lambda_\sigma}{\lambda_\rho}\right) = \lambda_\sigma^2\Big\langle\Big\langle \left(\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(s^0_{q_d})\right)^2 \Big\rangle\Big\rangle_{s^0_{q_d}} \tag{72}$$

$$\frac{\lambda_\rho}{\lambda_\sigma} - 1 = \lambda_\sigma\Big\langle\Big\langle \mathcal{M}_{\lambda_\sigma}[\,\sigma\,]''(s^0_{q_d}) \Big\rangle\Big\rangle \tag{73}$$

Derivation.—

In Appendix C.1 we derive the relation between the proximal map and Moreau envelope:

$$\mathcal{P}_\lambda[\,f\,](x) = x - \lambda\mathcal{M}_\lambda[\,f\,]'(x). \tag{74}$$

By applying the above relation to (59) and (60) respectively we arrive at (70) and (71). Next we derive (72), starting with (62):

$$\lambda_\rho = \lambda_\sigma\Big\langle\Big\langle \mathcal{P}_{\lambda_\sigma}[\,\sigma\,]'(s^0_{q_d}) \Big\rangle\Big\rangle = \lambda_\sigma\left(1 - \Big\langle\Big\langle \lambda_\sigma\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]''(s^0_{q_d}) \Big\rangle\Big\rangle\right). \tag{75}$$

Rearranging yields

$$1 - \frac{\lambda_\rho}{\lambda_\sigma} = \lambda_\sigma\Big\langle\Big\langle \mathcal{M}_{\lambda_\sigma}[\,\sigma\,]''(s^0_{q_d}) \Big\rangle\Big\rangle. \tag{76}$$

Finally, to derive (73), we begin with (61) to obtain

$$q_s = \Big\langle\Big\langle \left(\mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0 + \sqrt{q_d}z) - s^0\right)^2 \Big\rangle\Big\rangle_{s^0,z} = \Big\langle\Big\langle \left(\sqrt{q_d}z - \lambda_\sigma\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(s^0 + \sqrt{q_d}z)\right)^2 \Big\rangle\Big\rangle_{s^0,z}. \tag{77}$$

By expanding the square above we find

$$q_s = q_d - 2\sqrt{q_d}\Big\langle\Big\langle z\lambda_\sigma\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(s^0 + \sqrt{q_d}z) \Big\rangle\Big\rangle_{s^0,z} + \lambda_\sigma^2\Big\langle\Big\langle \left(\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(s^0_{q_d})\right)^2 \Big\rangle\Big\rangle_{s^0_{q_d}}. \tag{78}$$

Now, we again use the fact that for a gaussian random variable z, $\langle\langle\, zf(z) \,\rangle\rangle_z = \langle\langle\, f'(z) \,\rangle\rangle_z$, to show that

$$\sqrt{q_d}\Big\langle\Big\langle z\lambda_\sigma\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(s^0 + \sqrt{q_d}z) \Big\rangle\Big\rangle_{s^0,z} = q_d\Big\langle\Big\langle \lambda_\sigma\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]''(s^0_{q_d}) \Big\rangle\Big\rangle_{s^0_{q_d}} = q_d\left(1 - \frac{\lambda_\rho}{\lambda_\sigma}\right), \tag{79}$$

where the last equality follows from (73). Substituting this into (78) yields

$$q_s = q_d - 2q_d\left(1 - \frac{\lambda_\sigma}{\lambda_\rho}\right) + \lambda_\sigma^2\Big\langle\Big\langle \left(\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(s^0_{q_d})\right)2 \Big\rangle\Big\rangle_{s^0_{q_d}}, \tag{80}$$

which may be rearranged as

$$q_s + q_d\left(1 - 2\frac{\lambda_\sigma}{\lambda_\rho}\right) = \lambda_\sigma^2\Big\langle\Big\langle \left(\mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(s^0_{q_d})\right)^2 \Big\rangle\Big\rangle_{s^0_{q_d}}. \tag{81}$$

Finally, note that dividing (70) by the square of (71) yields the form of the replica equation for q_d exhibited in the main paper.

2.4 Accuracy of inference from the perspective of noise

Just as q_s measures inference performance through the MSE (Mean Squared Error) in estimating the signal, it is also useful to capture inference performance through the MSE in estimating the noise, which we work out in this section. This measure of inference performance, from the perspective of noise, will yield insights into the nature of optimal inference, as well as the nature of overfitting and generalization (see sections 5.1 and 5.2 below).

Result 6. *We define* $q_\epsilon = \left\langle\!\left\langle \left(\hat{\epsilon}(\epsilon_{q_s}) - \epsilon\right)^2 \right\rangle\!\right\rangle$, *which corresponds to replica prediction for the MSE of an M-estimator in estimating the the noise* ϵ. *In the regularized case we find:*

$$q_\epsilon = q_s + \frac{q_s}{\alpha}\left[\frac{q_d}{q_s}\left(\frac{\lambda_\rho}{\lambda_\sigma} - \frac{q_s}{q_d}\right)^2 - \frac{q_s}{q_d}\right]. \tag{82}$$

In the unregularized case and the over-sampled regime ($\alpha > 1$), $q_d = q_s$ *and* $\lambda_\rho = \lambda_\sigma$, *so this reduces to*

$$q_\epsilon = q_s\left(\frac{\alpha - 1}{\alpha}\right). \tag{83}$$

Derivation.—
Using the definition of q_ϵ and the relationship between the Moreau envelope and proximal map (74), we obtain

$$q_\epsilon = \left\langle\!\left\langle (\hat{\epsilon}(\epsilon_{q_s}) - \epsilon)^2 \right\rangle\!\right\rangle = \left\langle\!\left\langle (\hat{\epsilon}(\epsilon_{q_s}) - \epsilon_{q_s} + \sqrt{q_s} z)^2 \right\rangle\!\right\rangle = \left\langle\!\left\langle \left(\sqrt{q_s} z - \lambda_\rho \mathcal{M}_{\lambda_\rho}[\,\rho\,]'(\epsilon_{q_s})\right)^2 \right\rangle\!\right\rangle. \tag{84}$$

Expanding the square above yields

$$q_\epsilon = q_s - 2\sqrt{q_s}\lambda_\rho \left\langle\!\left\langle z \mathcal{M}_{\lambda_\rho}[\,\rho\,]'(\epsilon_{q_s}) \right\rangle\!\right\rangle + \lambda_\rho^2 \left\langle\!\left\langle \left(\mathcal{M}_{\lambda_\rho}[\,\rho\,]'(\epsilon_{q_s})\right)^2 \right\rangle\!\right\rangle. \tag{85}$$

Applying the identity $\langle\, z f(z)\,\rangle_z = \langle\langle\, f'(z)\,\rangle\rangle_z$ where z is a unit gaussian random variable z, yields

$$q_\epsilon = q_s - 2q_s \left\langle\!\left\langle \lambda_\rho \mathcal{M}_{\lambda_\rho}[\,\rho\,]''(\epsilon_{q_s}) \right\rangle\!\right\rangle + \left\langle\!\left\langle \left(\lambda_\rho \mathcal{M}_{\lambda_\rho}[\,\rho\,]'(\epsilon_{q_s})\right)^2 \right\rangle\!\right\rangle. \tag{86}$$

Then, substituting the replica symmetry equations (70),(71), yields

$$q_\epsilon = q_s - 2\frac{q_s\lambda_\rho}{\alpha\lambda_\sigma} + \frac{q_d\lambda_\rho^2}{\alpha\lambda_\sigma^2} = q_s + \frac{q_s}{\alpha}\left[\frac{q_d}{q_s}\left(\frac{\lambda_\rho}{\lambda_\sigma} - \frac{q_s}{q_d}\right)^2 - \frac{q_s}{q_d}\right]. \tag{87}$$

2.5 Noise-free replica equations

Motivated by the theory of compressed sensing which yields the interesting result that we can sometimes exactly recover unknown sparse signals in the undersampled measurement regime $\alpha < 1$ in the absence of noise via L_1 minimization, we consider the more general case of arbitrary convex inference in the absence of noise.

Result 7. *In the limit of noise-free measurements,* $y_\mu = \mathbf{X}_\mu \cdot s^0$. *The following equalities hold in the regime where the MSE error is non-zero*

$$\hat{s} = \min_s \sum_{j=1}^{P} \sigma(s_j) \qquad s.t \qquad \mathbf{y} = \mathbf{X}\mathbf{s}. \tag{88}$$

$$q_d = \frac{1}{\alpha} q_s, \tag{89}$$

$$\lambda_\rho = \alpha\lambda_\sigma, \tag{90}$$

$$q_s = \left\langle\!\left\langle (\hat{s}(s^0_{q_d}) - s^0)^2 \right\rangle\!\right\rangle_{s^0, z_s}, \tag{91}$$

$$\left\langle\!\left\langle \hat{s}'(s^0_{q_d}) \right\rangle\!\right\rangle_{s^0_{q_d}} = \alpha, \tag{92}$$

$$\textit{where} \qquad \hat{s}(s^0_{q_d}) = \mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0_{q_d}). \tag{93}$$

Derivation.—
The problem is equivalent to a M-estimator of the form

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \left[\sum_\mu \frac{\gamma}{2}\left(y_\mu - X_{\mu j} s_j\right)^2 + \sum_j \sigma(s_j)\right], \tag{94}$$

where we will take the limit $\gamma \to \infty$ to strictly enforce equality on the constraint equations. Before we take this limit, we substitute the form of ρ into Result 3, yielding the following 4 equations:

$$\frac{q_d}{\lambda_\sigma^2} = \frac{\alpha q_s \gamma^2}{(1+\gamma\lambda_\rho)^2} \tag{95}$$

$$\frac{1}{\lambda_\sigma} = \frac{\alpha\gamma}{1+\gamma\lambda_\rho} \tag{96}$$

$$q_s = \left\langle\!\left\langle \left(\mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0_{q_d}) - s^0\right)^2 \right\rangle\!\right\rangle_{s^0, z_s} \tag{97}$$

$$\lambda_\rho = \lambda_\sigma \left\langle\!\left\langle \mathcal{P}_{\lambda_\sigma}[\,\sigma\,]'(s^0_{q_d}) \right\rangle\!\right\rangle_{s^0_{q_d}}. \tag{98}$$

One can substitute (96) into (95) to show

$$q_s = \alpha q_d. \tag{99}$$

Now we consider two cases as $\gamma \to \infty$. First, in the finite error regime, q_s remains $O(1)$ as $\gamma \to \infty$, $\gamma\lambda_\rho \to \infty$. Second, in the error-free regime $q_s \to 0$, while $\gamma\lambda_\rho$ remains $O(1)$. In the finite error regime, the equations simplify and (96) becomes $\lambda_\rho = \alpha\lambda_\sigma$. Then (98) becomes

$$\left\langle\!\left\langle \mathcal{P}_{\lambda_\sigma}[\,\sigma\,]'(s^0_{q_d}) \right\rangle\!\right\rangle_{s^0_{q_d}} = \alpha. \tag{100}$$

Thus we obtain all the results in the box above corresponding to the finite error regime. We note that in the zero error limit, (96) becomes

$$\lambda_\sigma = \frac{1+\gamma\lambda_\rho}{\alpha\gamma} \to 0, \tag{101}$$

since $\gamma\lambda_\rho = O(1)$ as $\gamma \to \infty$. Applying the fact that $\lambda_\sigma = 0$ to (97) yields $q_s = q_d$, and to (98) yields $\lambda_\rho = \lambda_\sigma = 0$. Thus we find that in the zero error regime, as expected $\lambda_\rho, \lambda_\sigma, q_d, q_s \to 0$ is a solution to the RS equations. See [3] for a more detailed study of noise-free perfect inference for the specific case of L_1 minimization.

3 Analytic example: quadratic loss and regularization

3.1 General quadratic MSE

Result 8. *As an example, we derive the asymptotic MSE, q_s, analytically for the case of M-estimators with a square loss function and quadratic regularization:*

$$\rho(x) = \frac{x^2}{2} \qquad\qquad \sigma(x) = \gamma\frac{x^2}{2}, \tag{102}$$

we find the normalized MSE (i.e. the fraction of unexplained variance) to be

$$\bar{q}_s = \frac{q_s}{\left\langle\!\left\langle (s^0)^2 \right\rangle\!\right\rangle} = \frac{\phi + \alpha\gamma^2\lambda_\sigma^2}{\alpha(1+\gamma\lambda_\sigma)^2 - 1}, \tag{103}$$

where $\phi = \dfrac{\left\langle\!\left\langle \epsilon^2 \right\rangle\!\right\rangle}{\left\langle\!\left\langle (s^0)^2 \right\rangle\!\right\rangle} = \frac{1}{SNR}$ is the inverse signal to noise ratio, and λ_σ is the non-negative root which solves the quadratic equation:

$$\alpha\gamma\lambda_\sigma^2 + (\alpha - \gamma - 1)\lambda_\sigma - 1 = 0. \tag{104}$$

Derivation. —
From Result 3, the fact that $f(x) = \frac{\gamma}{2}x^2$, $\mathcal{P}_\lambda[\,f\,](x) = \frac{x}{1+\lambda\gamma}$, allows us to rewrite (63) and (64) as

$$\hat{\epsilon}(\epsilon_{q_s}) = \mathcal{P}_{\lambda_\rho}[\,\rho\,](\epsilon_{q_s}) = \frac{\epsilon_{q_s}}{1+\lambda_\rho}, \tag{105}$$

$$\hat{s}(s^0_{q_d}) = \mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0_{q_d}) = \frac{s^0_{q_d}}{1+\gamma\lambda_\sigma}. \tag{106}$$

Substituting the form of $\hat{\epsilon}, \hat{s}$ into (59 - 62) yields

$$\frac{q_d}{\lambda_\sigma^2} = \frac{\alpha}{\lambda_\rho^2} = \frac{\alpha}{(1+\lambda_\rho)^2}\langle\!\langle\, \epsilon_{q_s}^2 \,\rangle\!\rangle_{\epsilon_{q_s}} \tag{107}$$

$$\frac{1}{\lambda_\sigma} = \frac{\alpha}{1+\lambda_\rho} \implies \lambda_\sigma = \frac{1+\lambda_\rho}{\alpha} \tag{108}$$

Combining (108) and (107) yields

$$q_d = \frac{\langle\!\langle\, \epsilon_{q_s}^2 \,\rangle\!\rangle}{\alpha} = \frac{\langle\!\langle\, \epsilon^2 \,\rangle\!\rangle + q_s}{\alpha} \tag{109}$$

$$q_s = \left\langle\!\!\left\langle \left(\frac{s^0 + z\sqrt{q_d}}{1+\lambda_\sigma\gamma} - s^0 \right)^2 \right\rangle\!\!\right\rangle_{s^0,z}, \tag{110}$$

$$\lambda_\rho = \frac{\lambda_\sigma}{1+\gamma\lambda_\sigma} \tag{111}$$

which implies

$$q_s = \frac{q_d + (\lambda_\sigma\gamma)^2 \langle\!\langle\, (s^0)^2 \,\rangle\!\rangle_{s^0}}{(1+\lambda_\sigma\gamma)^2}. \tag{112}$$

Combining (108) and (111) we find that λ_σ must satisfy the quadratic equation

$$\alpha\gamma\lambda_\sigma^2 + (\alpha - \gamma - 1)\lambda_\sigma - 1 = 0. \tag{113}$$

The constraint of non-negativity forces λ_σ to be the positive root, which is unique. Finally, substituting (109) into (112) yields

$$q_s = \frac{\langle\!\langle\, \epsilon^2 \,\rangle\!\rangle + q_s + \alpha\,(\gamma\lambda_\sigma)^2 \,\langle\!\langle\, (s^0)^2 \,\rangle\!\rangle}{\alpha\,(1+\gamma\lambda_\sigma)^2}. \tag{114}$$

Re-arranging to solve for q_s and dividing by the variance of the parameters, we find

$$\bar{q}_s = \frac{q_s}{\langle\!\langle\, (s^0)^2 \,\rangle\!\rangle} = \frac{\phi + \alpha(\gamma\lambda_\sigma)^2}{\alpha(1+\gamma\lambda_\sigma)^2 - 1}, \tag{115}$$

where we define $\phi = \dfrac{\langle\!\langle\, \epsilon^2 \,\rangle\!\rangle}{\langle\!\langle\, (s^0)^2 \,\rangle\!\rangle}$ to be the noise to signal ratio. Note that in the limit of strong regularization ($\gamma \to \infty$), this reduces to $q_s = \langle\!\langle\, (s^0)^2 \,\rangle\!\rangle$ which is intuitive since the strong ridge penalty pins the estimator $\hat{s}$ to 0. The other extreme is the un-regularized M-estimator, which corresponds simply to least squares regression and yields

$$q_s = \frac{\langle\!\langle\, \epsilon^2 \,\rangle\!\rangle}{\alpha - 1}, \tag{116}$$

which demonstrates that the MSE is unbounded as the number of samples approaches the number of dimensions.

3.2 Optimal quadratic MSE

Result 9. ***Optimal Quadratic M-estimation****: We find that the optimal quadratic M-estimator has the form:*

$$\rho(x) = \frac{x^2}{2} \qquad\qquad \sigma(x) = \phi\frac{x^2}{2}, \tag{117}$$

and satisfies

$$\bar{q}_s = \frac{1 - \alpha - \phi + \sqrt{(\phi + \alpha - 1)^2 + 4\phi}}{2}. \tag{118}$$

This result does not depend on the general form of P_ϵ and P_s except via ϕ, the inverse SNR.

Derivation of Optimal Regularization Parameter:
To find the value of γ that minimizes $\bar{q}_s$, we set the derivative of (103) with respect to γ equal to zero:

$$\frac{d\bar{q}_s}{d\gamma} = \frac{2\alpha(\lambda_\sigma + \gamma\frac{d\lambda_\sigma}{d\gamma})}{\left(\alpha(1+\gamma\lambda_\sigma)^2 - 1\right)^2}\left(\alpha\gamma^2\lambda_\sigma^2 + (\alpha - \phi - 1)\gamma\lambda_\sigma - \phi\right) = 0. \tag{119}$$

We look for when

$$\alpha\gamma^2\lambda_\sigma^2 + (\alpha - \phi - 1)\gamma\lambda_\sigma - \phi = 0, \tag{120}$$

we then multiply (104) by γ:

$$\alpha\gamma^2\lambda_\sigma^2 = \gamma - (\alpha - \gamma - 1)\gamma\lambda_\sigma. \tag{121}$$

Substituting this result into (120) yields

$$(\gamma - \phi)(1 + \gamma\lambda_\sigma) = 0, \tag{122}$$

so we find a fixed point of $\bar{q}_s$ occurs at $\gamma = \phi$.

Derivation of Optimal $\bar{q}_s$
Substituting the optimal regularization parameter $\gamma = \phi$ into (104) and multiplying by an overall factor of ϕ yields

$$\alpha(\phi\lambda_\sigma)^2 + (\alpha - \phi - 1)\phi\lambda_\sigma - \phi = 0, \tag{123}$$

multiplying both sides by the positive factor $1 + \phi + \alpha$, yields

$$\alpha(1 + \phi + \alpha)(\phi\lambda_\sigma)^2 + (1 + \phi + \alpha)(\alpha - \phi - 1)\phi\lambda_\sigma - \phi(\phi + \alpha + 1) = 0. \tag{124}$$

This can be expanded an re-arranged as

$$\phi(\alpha\lambda_\sigma - 1)\left(\phi\lambda_\sigma(1 + \phi + \alpha) + \phi + \alpha - 1\right) = \phi\lambda_\sigma(1 + \phi - \alpha) + 2\phi. \tag{125}$$

It follows that

$$\phi(\alpha\lambda_\sigma - 1) = \frac{\phi\lambda_\sigma(1 + \phi - \alpha) + 2\phi}{\left(\phi\lambda_\sigma(1 + \phi + \alpha) + \phi + \alpha - 1\right)}. \tag{126}$$

Substitution of the the form of $(\phi\lambda_\sigma)^2$ from (123) into (103) yields

$$\bar{q}_s = \frac{\phi + \alpha\phi^2\lambda_\sigma^2}{\alpha(1 + \phi\lambda_\sigma)^2 - 1} = \frac{\phi\lambda_\sigma(1 + \phi - \alpha) + 2\phi}{\left(\phi\lambda_\sigma(1 + \phi + \alpha) + \phi + \alpha - 1\right)}. \tag{127}$$

Combining the previous two equations implies that

$$\bar{q}_s = \phi(\alpha\lambda_\sigma - 1). \tag{128}$$

We complete the derivation by substituting the value of $\phi\lambda_\sigma$, which is the positive root of a quadratic equation (123):

$$\phi\lambda_\sigma = \frac{\phi + 1 - \alpha + \sqrt{(\phi + 1 - \alpha)^2 + 4\alpha\phi}}{2\alpha}. \tag{129}$$

Substituting (129) into (128) yields

$$\bar{q}_s = \frac{\phi + 1 - \alpha + \sqrt{(\phi + 1 - \alpha)^2 + 4\alpha\phi} - 2\phi}{2} = \frac{1 - \alpha - \phi + \sqrt{(\phi + \alpha - 1)^2 + 4\phi}}{2}. \tag{130}$$

3.3 Training and test errors

$$\bar{\mathcal{E}}^{\text{test}} = \frac{\mathcal{E}^{\text{test}}}{\left\langle\!\left\langle (s^0)^2 \right\rangle\!\right\rangle} = \bar{q}_s + \phi, \tag{131}$$

$$\mathcal{E}^{\text{train}} = \left\langle\!\left\langle \mathcal{P}_{\lambda_\rho}[\,\rho\,](\epsilon_{q_s})^2 \right\rangle\!\right\rangle. \tag{132}$$

Under our choice of $\rho(x) = \frac{x^2}{2}$, it follows that

$$\mathcal{E}^{\text{train}} = \frac{\left\langle\!\left\langle \epsilon^2 \right\rangle\!\right\rangle + q_s}{(1 + \lambda_\rho)^2}. \tag{133}$$

Combining (108) and (111) to find a quadratic equation in λ_ρ and assuming the optimal setting $\gamma = \phi$ yields

$$(\phi\lambda_\rho)^2 + (\phi\lambda_\rho)(\alpha + \phi - 1) - \phi = 0. \tag{134}$$

This implies that in the optimal quadratic setting

$$\lambda_\rho = \frac{\bar{q}_s}{\phi}, \tag{135}$$

solving for $\bar{\mathcal{E}}^{\text{train}} = \frac{\mathcal{E}^{\text{train}}}{\left\langle\!\left\langle (s^0)^2 \right\rangle\!\right\rangle}$ yields

$$\bar{\mathcal{E}}^{\text{train}} = \frac{\phi^2}{\bar{\mathcal{E}}^{\text{test}}}. \tag{136}$$

3.4 Limiting cases and scaling

Result 10. ***Limiting case for optimal quadratic inference:***

In the limit SNR $\gg 1$*:*

$$\bar{q}_s = \begin{cases} 1-\alpha & \alpha < 1, \\ \frac{1}{\sqrt{SNR}} & \alpha = 1 \\ \frac{1}{SNR(\alpha-1)} & \alpha > 1 \end{cases} \tag{137}$$

Finite SNR:

$$\bar{q}_s = \begin{cases} 1-\alpha\frac{SNR}{SNR+1} & \alpha \ll 1, \\ \frac{1}{2SNR}\left(\sqrt{1+4SNR}-1\right) & \alpha = 1 \\ \frac{1}{SNR(\alpha-1)+1} & \alpha \gg 1 \end{cases} \tag{138}$$

In the limit SNR $\ll 1$

$$\bar{q}_s = \frac{1}{SNR(\alpha-1)+1} \tag{139}$$

The following form will be convenient for expanding $\bar{q}_s$ in different limits. Factoring $\alpha+\phi-1$ out of (118) yields,

$$\bar{q}_s = \left(\frac{\alpha+\phi-1}{2}\right)\left(-1+\frac{\alpha+\phi-1}{|\alpha+\phi-1|}\sqrt{1+\frac{4\phi}{(\alpha+\phi-1)^2}}\right). \tag{140}$$

SNR $\gg$ 1:
First consider $\alpha = 1$, in which case (140) reduces to

$$\bar{q}_s = \frac{\phi}{2}\left(-1+\sqrt{1+\frac{4}{\phi}}\right) = \frac{1}{2\text{SNR}}\left(\sqrt{1+4\text{SNR}}-1\right). \tag{141}$$

In order to Taylor expand this in small ϕ, we write the above expression as

$$\bar{q}_s = \frac{\phi}{2}\left(-1+\left(\frac{4}{\phi}\right)^{1/2}\sqrt{1+\frac{\phi}{4}}\right) \approx \frac{\phi}{2}\left(-1+\left(\frac{4}{\phi}\right)^{1/2}\left(1+\frac{\phi}{8}\right)\right), \tag{142}$$

which approaches $\bar{q}_s = \sqrt{\phi} = \frac{1}{\sqrt{\text{SNR}}}$ in the large SNR limit. Next, in the case $\alpha < 1$, we expand (140) in small ϕ which yields

$$\bar{q}_s = \left(\frac{\alpha+\phi-1}{2}\right)\left(-1-\sqrt{1+\frac{4\phi}{(\alpha+\phi-1)^2}}\right) \approx 1-\alpha. \tag{143}$$

For $\alpha > 1$, the sign inside the square root changes, and expanding in small ϕ yields

$$\bar{q}_s = \left(\frac{\alpha+\phi-1}{2}\right)\left(-1+\sqrt{1+\frac{4\phi}{(\alpha+\phi-1)^2}}\right) \approx \frac{\phi}{\alpha+\phi-1} \to \frac{\phi}{\alpha-1}. \tag{144}$$

Finite SNR :
Consider the limit of finite SNR with a very low sampling rate $\alpha \ll 1$, in which case (140) may be written as

$$\bar{q}_s = \frac{1-\alpha-\phi+\sqrt{(\phi+1)^2+2\alpha(\phi-1)+\alpha^2}}{2} = \frac{1-\alpha-\phi+(1+\phi)\sqrt{1+\frac{2\alpha(\phi-1)+\alpha^2}{(1+\phi)^2}}}{2}. \tag{145}$$

Expanding this expression to first order in α yields

$$\bar{q}_s = \frac{1-\alpha-\phi+(1+\phi)\left(1+\frac{\alpha(\phi-1)}{(1+\phi)^2}\right)}{2} = \frac{2-\alpha+\alpha\frac{(\phi-1)}{1+\phi}}{2} = 1-\frac{\alpha}{1+\phi} = 1-\alpha\frac{\text{SNR}}{\text{SNR}+1}. \tag{146}$$

These results characterize the way that, with greater SNR, increasing the measurement density yields a greater reduction in the estimation error. The case of $\alpha = 1$ is already shown in (141). In the limit of $\alpha \gg 1$ with finite SNR:

$$\bar{q}_s = \left(\frac{\alpha+\phi-1}{2}\right)\left(-1+\sqrt{1+\frac{4\phi}{(\alpha+\phi-1)^2}}\right) \approx \left(\frac{\alpha+\phi-1}{2}\right)\left(-1+\left(1+2\frac{\phi}{(\alpha+\phi-1)^2}\right)\right) = \frac{1}{\text{SNR}(\alpha-1)+1}. \tag{147}$$

SNR $\ll$ 1:
In the case of very low SNR ($\phi \gg 1$) with finite α, the first order expansion of the expression (140) in small $\frac{1}{\phi}$ is

$$\bar{q}_s = \left(\frac{\alpha+\phi-1}{2}\right)\left(-1+\sqrt{1+\frac{4\phi}{(\alpha+\phi-1)^2}}\right) \approx \frac{\phi}{\alpha+\phi-1} = \frac{1}{\mathrm{SNR}(\alpha-1)+1}. \tag{148}$$

4 Lower bounds on the error of any convex inference procedure

4.1 Unregularized inference

Result 11. ***Unregularized High Dimensional Error Bound*** *In the over-sampled ($N > P$), but still high dimensional regime $\alpha > 1$, the asymptotic lower bound q_s (estimator MSE) for any unregularized convex M-estimator is given by:*

$$q_s \geq \frac{1}{\alpha J\left[\epsilon+\sqrt{q_s}z\right]} \geq \frac{1}{(\alpha-1)J\left[\,\epsilon\,\right]}. \tag{149}$$

The form of the Cramer-Rao bound in this problem is derived in Appendix B.5 to be:

$$q_s \geq \frac{1}{\alpha J\left[\,\epsilon\,\right]}. \tag{150}$$

Thus, the Cramer-Rao bound is not saturated for finite α. See Appendix B for more information on the Fisher information and Cramer-Rao bound.

We define f_{q_s} to be the pdf of $\epsilon+\sqrt{q_s}z$ where z is a unit gaussian random variable. The bound above follows from constraints (65) and (66) derived in Result 4, which may be written as

$$\int f_{q_s}(y) r'(y) dy = \frac{1}{\alpha}, \tag{151}$$

$$\int f_{q_s}(y) r(y)^2 dy = \frac{q_s}{\alpha}, \tag{152}$$

where we have defined

$$r(y) = \lambda_\rho \mathcal{M}_{\lambda_\rho}\left[\,\rho\,\right]'(y). \tag{153}$$

Integrating (151) by parts and squaring the result yields

$$\left(\frac{1}{\alpha}\right)^2 = \left(\int dy r(y) f'_{q_s}(y)\right)^2. \tag{154}$$

By breaking the integrand into two parts and applying the Cauchy-Schwartz inequality, we find:

$$\left(\frac{1}{\alpha}\right)^2 = \left(\int dy \frac{f'_{q_s}(y)}{\sqrt{f_{q_s}(y)}}\sqrt{f_{q_s}(y)} r(y)\right)^2 \leq J\left[\epsilon+\sqrt{q_s}z\right] \int f_{q_s}(y) r(y)^2 dy. \tag{155}$$

Substituting (152) into this yields the inequality

$$q_s \geq \frac{1}{\alpha J\left[\epsilon+\sqrt{q_s}z\right]}. \tag{156}$$

To complete the derivation, we apply the convolutional Fisher inequality (see Appendix B.2)

$$J\left[\,\epsilon_1+\epsilon_2\,\right] \leq \frac{J\left[\,\epsilon_1\,\right] J\left[\,\epsilon_2\,\right]}{J\left[\,\epsilon_1\,\right]+J\left[\,\epsilon_2\,\right]}, \tag{157}$$

where ϵ_1, ϵ_2 are random variables drawn from different distributions. This result implies

$$J\left[\epsilon+\sqrt{q_s}z\right] \leq \frac{J\left[\,\epsilon\,\right]}{1+q_s J\left[\,\epsilon\,\right]}. \tag{158}$$

Substituting this into (156) yields

$$q_s \geq \frac{1}{\alpha J\left[\epsilon+\sqrt{q_s}z\right]} \geq \frac{1+q_s J\left[\,\epsilon\,\right]}{\alpha J\left[\,\epsilon\,\right]}. \tag{159}$$

This implies the inequality $q_s J\left[\,\epsilon\,\right] \geq \frac{1}{\alpha}\left(1+q_s J\left[\,\epsilon\,\right]\right)$, which may be applied repeatedly to the RHS of (159) to show

$$q_s \geq \frac{1}{\alpha J\left[\epsilon+\sqrt{q_s}z\right]} \geq \frac{1}{\alpha J\left[\,\epsilon\,\right]}\left(1+\frac{1}{\alpha}+\frac{1}{\alpha^2}+\ldots\right). \tag{160}$$

For $\alpha > 1$, we can write this as

$$q_s \geq \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s} z\right]} \geq \frac{1}{(\alpha - 1) J\left[\,\epsilon\,\right]}. \tag{161}$$

4.2 Regularized inference

Result 12. ***High Dimensional Regularized Bounds***
For regularized M-estimators (regularizer is not assumed to be zero), we find that the estimator MSE q_s is lower bounded by

$$q_s \geq q_d - q_d^2 J\left[s^0 + \sqrt{q_d} z\right] = q_s^{MMSE}(q_d), \qquad \textit{where} \qquad q_d = \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s} z\right]}. \tag{162}$$

Here $q_s^{MMSE}(q_d)$ denotes the scalar minimum mean squared error of reconstructing random variable s^0 corrupted by gaussian noise with variance q_d via Bayesian inference of the posterior mean:

$$q_s^{MMSE}(q_d) = \left\langle\!\left\langle \left(s^0 - \left\langle\, s | s^0 + \sqrt{q_d} z\,\right\rangle\right)^2 \right\rangle\!\right\rangle_{s^0, z}. \tag{163}$$

The above bound also leads to the inequality:

$$q_s \geq \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s} z\right] + J\left[\,s^0\,\right]}, \tag{164}$$

which implies that regularized M-estimators can leverage a highly informative signal distribution to improve on the Cramer-Rao bound in the high dimensional setting.

In this result, we derive a lower bound on MSE q_s over all convex Regularized M-estimators defined by pairs (ρ, σ), given the constraints imposed by Result 5. For notational convenience, we define:

$$r_\rho\,(y) = \lambda_\rho \mathcal{M}_{\lambda_\rho}[\,\rho\,]'(y), \tag{165}$$

$$r_\sigma\,(y) = \lambda_\sigma \mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(y). \tag{166}$$

We also find it convenient to define f_{q_s}, g_{q_d} to be the pdf's for $\epsilon + \sqrt{q_s} z_\epsilon$ and $s^0 + \sqrt{q_d} z_s$ respectively where z_ϵ, z_s are independent unit gaussian random variables. Thus, (72) and (73) may be written as

$$\frac{\lambda_\rho}{\lambda_\sigma} - 1 = \int g'_{q_d}\,(y) r_\sigma\,(y) dy, \tag{167}$$

$$\int g_{q_d}\,(y) r_\sigma\,(y)^2 dy = q_s - q_d - 2q_d \int g'_{q_d}\,(y) r_\sigma\,(y) dy = q_s - q_d - 2q_d \left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right). \tag{168}$$

Squaring (167) and applying the Cauchy-Schwartz inequality yields

$$\left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right)^2 = \left(\int \frac{g'_{q_d}\,(y)}{\sqrt{g_{q_d}\,(y)}} \sqrt{g_{q_d}\,(y)} r_\sigma\,(y) dy\right)^2 \leq J\left[s^0 + \sqrt{q_d} z\right] \int g_{q_d}\,(y) r_\sigma\,(y)^2 dy. \tag{169}$$

Substituting (168) into the RHS of the above inequality yields

$$\left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right)^2 \leq J\left[s^0 + \sqrt{q_d} z\right] \left(q_s - q_d - 2q_d \left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right)\right). \tag{170}$$

Completing the square, we find

$$\left(\frac{\lambda_\rho}{\lambda_\sigma} - 1 + q_d J\left[s^0 + \sqrt{q_d} z\right]\right)^2 \leq J\left[s^0 + \sqrt{q_d} z\right] \left(q_s - q_d + q_d^2 J\left[s^0 + \sqrt{q_d} z\right]\right). \tag{171}$$

Since the LHS and Fisher information are non-negative, this implies (162):

$$q_s \geq q_d - q_d^2 J\left[s^0 + \sqrt{q_d} z\right]. \tag{172}$$

To derive the form of q_d in (162), we use the two constraint equations containing r_ρ (60) and (59). These may be written, by integrating (244) and (243) by parts, as

$$\int f'_{q_s}\,(y) r_\rho\,(y) dy = -\frac{\lambda_\rho}{\alpha \lambda_\sigma}, \tag{173}$$

$$\int f_{q_s}\,(y) r_\rho\,(y)^2 dy = \frac{q_d \lambda_\rho^2}{\alpha \lambda_\sigma^2}. \tag{174}$$

Squaring (173) and applying Cauchy Schwartz as before, yields

$$\left(\frac{\lambda_\rho}{\alpha\lambda_\sigma}\right)^2 = \left(\int f'_{q_s}(y) r_\rho(y) dy\right)^2 \leq J\left[\epsilon + \sqrt{q_s} z\right] \int f_{q_s}(y) r_\rho(y)^2 dy. \tag{175}$$

By inserting (174) into the RHS, we find

$$q_d \geq \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s} z\right]}. \tag{176}$$

By applying the relation between MMSE and Fisher information in Appendix B.4 to (172), we find that q_s is lower bounded by the MMSE error of reconstruction of s^0 given gaussian noise of variance q_d:

$$q_s \geq q_s^{\mathrm{MMSE}}(q_d). \tag{177}$$

Note that the function on the RHS monotonically increasing in q_d since increasing the variance in the noise will make reconstruction more difficult. It follows that the strongest bound occurs for the largest possible value of q_d:

$$q_d = \frac{1}{\alpha J\left[\, \epsilon + \sqrt{q_s} z \,\right]} \tag{178}$$

To complete the derivation, we apply the Fisher information convolution inequality (157) (see Appendix B.2) which implies that

$$J\left[s^0 + \sqrt{q_d} z\right] \leq \frac{J\left[\, s^0 \,\right]}{1 + q_d J\left[\, s^0 \,\right]}, \tag{179}$$

substituting this into (172) yields

$$q_s \geq q_d - q_d^2 J\left[s^0 + \sqrt{q_d} z\right] \geq \frac{q_d}{1 + q_d J\left[\, s^0 \,\right]}, \tag{180}$$

and since the RHS is monotonically increasing in q_d, it is lower bounded by the lowest allowed value of q_d which is bounded by (176) so that

$$q_s \geq \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s} z\right] + J\left[\, s^0 \,\right]}. \tag{181}$$

Result 13. ***Noise-free inference***
One application of the high dimensional regularized bound above is that it produces a lower bound on the required density of samples α_r to achieve some permissible mean squared error per parameter Q in the big data limit only in terms of the allowed error Q and the Fisher information of the noise $J\left[\,\epsilon\,\right]$ and parameters $J\left[\, s^0 \,\right]$.

$$\alpha_r \geq \frac{1 + QJ\left[\,\epsilon\,\right]}{J\left[\,\epsilon\,\right]}\left(\frac{1}{Q} - J\left[\, s^0 \,\right]\right). \tag{182}$$

Derivation.—
Re-arranging (164) we find

$$\alpha \geq \frac{1 - q_s J\left[\, s^0 \,\right]}{q_s J\left[\epsilon + \sqrt{q_s} z\right]} \geq \frac{(1 - q_s J\left[\, s^0 \,\right])(1 + q_s J\left[\,\epsilon\,\right])}{q_s J\left[\,\epsilon\,\right]}, \tag{183}$$

where the second inequality follows from (157). Differentiating the RHS of (183) with respect to q_s, we find that it is a monotonically decreasing quantity since the Fisher information is positive. Thus, under the condition $q_s \leq Q$ the strongest bound occurs at $q_s = Q$ and is

$$\alpha_r \geq \max_{q_s \leq Q}\left[\frac{(1 - q_s J\left[\, s^0 \,\right])(1 + q_s J\left[\,\epsilon\,\right])}{q_s J\left[\,\epsilon\,\right]}\right] = \left[\frac{(1 - QJ\left[\, s^0 \,\right])(1 + QJ\left[\,\epsilon\,\right])}{QJ\left[\,\epsilon\,\right]}\right]. \tag{184}$$

Note that this inequality reveals how both the Fisher information of the noise and signal contributed to a lower bound on the requisite measurement density to achieve a prescribed error.

4.3 Lower bound on noise-free inference

Result 14. *In the noise free regime ($\epsilon \to 0$), we find the following lower bound on the MSE q_s*

$$q_s \geq q_s^{MMSE}(q_d), \quad \textit{where} \quad q_d = \frac{q_s}{\alpha}, \tag{185}$$

and in terms of Fisher information,

$$q_s \geq \frac{\alpha(1-\alpha)}{J\left[\, s^0 + \sqrt{\frac{q_s}{\alpha}} z \,\right]} \geq \frac{1-\alpha}{J\left[\, s^0 \,\right]}. \tag{186}$$

Derivation.—
Note that the Result can be derived directly from the regularized bound (162), which simplifies in the noise-free limit $\epsilon \to 0$ so that $q_d = \frac{q_s}{\alpha}$ and

$$q_s \geq q_s^{\text{MMSE}}(q_d). \tag{187}$$

From the relation between scalar Fisher information and MMSE (see Appendix B.2), the above equation may be written as

$$q_s \geq q_d(1 - q_d J\left[\, s^0_{q_d} \,\right]) = \frac{q_s}{\alpha}(1 - \frac{q_s}{\alpha} J\left[\, s^0_{q_d} \,\right]). \tag{188}$$

By dividing both sides of this equation by q_s and rearranging, we arrive at the inequality

$$q_s \geq \frac{\alpha(1-\alpha)}{J\left[\, s^0_{q_d} \,\right]} = \frac{\alpha(1-\alpha)}{J\left[\, s^0 + \sqrt{\frac{q_s}{\alpha}} z \,\right]}. \tag{189}$$

From the convolutional Fisher inequality (see Appendix B.2) we know

$$\frac{1}{J\left[\, s^0_{q_d} \,\right]} \geq q_d + \frac{1}{J\left[\, s^0 \,\right]}, \tag{190}$$

substituting this into inequality into the previous equation, after some rearrangement, yields

$$q_s \geq \frac{1-\alpha}{J\left[\, s^0 \,\right]}. \tag{191}$$

4.4 Example: bounds on compressed sensing

Result 15. *For noise free compressed sensing, where the signal s^0 is drawn from a distribution which is f-sparse (only a fraction f of its components are nonzero), we find that the optimal convex M-estimator requires a number of measurements N satisfying*

$$N \geq fP. \tag{192}$$

to achieve perfect reconstruction to be possible in the asymptotic limit.

Consider the compressed sensing problem: the parameters being estimated are f-sparse so that a fraction f of them are non-zero. In this scenario, the bound from the previous result becomes

$$\lim_{q_s \to 0} J\left[\, s^0 + \sqrt{\frac{qs}{\alpha}} z \,\right] = \alpha \frac{1-f}{q_s} + O(1). \tag{193}$$

Substituting this into (189) we find that in the limit of perfect reconstruction, we have

$$\alpha \geq f. \tag{194}$$

In other words we must have,

$$\frac{N}{P} \geq f, \tag{195}$$

so the number of measurements required is greater than the size of signal multiplied by the fraction of non-zero elements for an M-estimator to achieve perfect reconstruction in the asymptotic limit.

5 Optimal inference

5.1 Unregularized inference

Result 16. *In the case of unregularized regression ($\sigma = 0$) and log-concave noise probability density P_ϵ (or equivalently convex noise energy $E_\epsilon = -\ln P_\epsilon$), we find that the loss function with minimal MSE is*

$$\rho^{opt} = -\mathcal{M}_{q_s}[\,\ln P_{\epsilon_{q_s}}\,]. \tag{196}$$

$\mathcal{M}_\lambda[\,f\,]$ is a Moreau envelope defined as

$$\mathcal{M}_\lambda[\,f\,](x) = \min_y \left[\frac{(x-y)^2}{2\lambda} + f(y)\right], \tag{197}$$

(see Appendix C.1) and q_s is the minimal solution satisfying

$$q_s = \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s} z\right]}. \tag{198}$$

The distribution of signal and noise estimates are obtained through

$$\hat{s} = \sqrt{q_s} z + s^0, \tag{199}$$

$$\hat{\epsilon} = \hat{\epsilon}^{MMSE}(\epsilon_{q_s}) = \int \epsilon P(\epsilon|\epsilon_{q_s}) d\epsilon, \tag{200}$$

with probabilities $P(z)$ and $P(\epsilon_{q_s})$ respectively where z is is a unit gaussian random variable. The MSE per component in estimating the noise is

$$q_\epsilon = q_\epsilon^{MMSE}(q_s) = \left\langle\!\!\left\langle \left(\left\langle\, \epsilon|\epsilon + \sqrt{q_s} z\right\rangle - \epsilon\right)^2 \right\rangle\!\!\right\rangle_{\epsilon,z}. \tag{201}$$

The training and test errors are

$$\mathcal{E}^{train} = \sigma_\epsilon^2 - q_\epsilon = \sigma_\epsilon^2 - q_\epsilon^{MMSE}(q_s), \qquad \mathcal{E}^{test} = \sigma_\epsilon^2 + q_s. \tag{202}$$

Derivation of the Algorithm
For ease of notation we will write $f = P_\epsilon$ and $\epsilon_q = \epsilon + \sqrt{q} z$, where z is a unit gaussian random variable, so that ϵ_q is a random variable drawn from pdf f_q. Our goal is to minimize the error q_s by optimizing ρ under the unregularized constraints from Result 4 using the method of lagrange multipliers and calculus of variations. Under the definition $r(x) = \lambda_\rho \mathcal{M}_{\lambda_\rho}[\,\rho\,]'(x)$, the unregularized constraints (68) and (69) are

$$\alpha \int f_{q_s}(y) r(y)^2 dy - q_s = 0, \tag{203}$$

$$\int f_{q_s}(y) r'(y) dy - \frac{1}{\alpha} = 0. \tag{204}$$

Introducing multipliers λ and μ, our lagrangian may be written

$$L = q_s + \lambda\left[\int f_{q_s}(y) r'(y) dy - \frac{1}{\alpha}\right] + \mu\left[\int f_{q_s}(y)\,(r(y))^2\,dy - \frac{q_s}{\alpha}\right]. \tag{205}$$

The strategy we employ is to first find the optimal r, and second to derive ρ^{opt} from r. Calculus of variations may be applied if we combine the integrals in the Lagrangian so that

$$L = q_s - \lambda\frac{1}{\alpha} - \mu\frac{1}{\alpha} q_s + \int S(r(y), r'(y)) dy, \tag{206}$$

where

$$S(r, r') = \left(\lambda r' + \mu r^2\right) f_{q_s}. \tag{207}$$

Defining δL to be the variational derivative of L (induced by varying r by a small change δr). Taylor expanding the action S in the integrand gives:

$$\delta L = \int \left(\frac{\partial S}{\partial r}\delta r + \frac{\partial S}{\partial r'}\delta r'\right) dy. \tag{208}$$

Integrating the second term by parts gives

$$\delta L = \int \left(\frac{\partial S}{\partial r} - \frac{d}{dy}\frac{\partial S}{\partial r'} \right) \delta r dy. \tag{209}$$

For stationarity to hold for all δr we thus require the Euler-Lagrange (E-L) equation to hold, i.e.

$$\frac{\partial S}{\partial r} = \frac{\partial}{\partial y}\frac{\partial S}{\partial r'}.. \tag{210}$$

Inserting (207) into (210) yields

$$2\mu r f_{q_s} = \lambda f'_{q_s}. \tag{211}$$

Re-arranging,

$$r(y) = \frac{d}{dy}\frac{\lambda}{2\mu}\ln(f_{q_s}(y)). \tag{212}$$

To solve for $\frac{\lambda}{2\mu}$, we extremize the Lagrangian with respect to other variables after substituting the form of (211) into (205) to find

$$L = q_s - \lambda\frac{1}{\alpha} - \frac{1}{\alpha}\mu q_s - \frac{\lambda^2}{4\mu}\int \frac{(f'_{q_s})^2}{f_{q_s}}. \tag{213}$$

We can now use the fact that $J\left[\epsilon + \sqrt{q_s}z\right] = \int \frac{(f'_{q_s})^2}{f_{q_s}}$ to simplify notation. Now, extremizing L with respect to λ gives $\frac{\lambda}{2\mu} = -\frac{1}{\alpha J\left[\epsilon+\sqrt{q_s}z\right]}$, and with respect to μ gives $\left(\frac{\lambda}{2\mu}\right)^2 = \frac{q_s}{\alpha J\left[\epsilon+\sqrt{q_s}z\right]}$. Combining these results we find $\frac{\lambda}{2\mu} = -q_s$. Note that the zero solution can be neglected since $\frac{\lambda}{2\mu} = \frac{1}{\alpha J\left[\epsilon+\sqrt{q_s}z\right]}$. Substituting into L gives

$$L = q_s + \frac{\mu}{\alpha J\left[\epsilon + \sqrt{q_s}z\right]}\left(\frac{1}{\alpha} - q_s J\left[\epsilon + \sqrt{q_s}z\right]\right). \tag{214}$$

Because μ is still a free parameter, we choose q_s to be the minimal value satisfying

$$J\left[\epsilon + \sqrt{q_s}z\right] q_s = \frac{1}{\alpha}. \tag{215}$$

Now we can substitute $\frac{\lambda}{2\mu} = -q_s$ into (212) yields

$$r(x) = -q_s\frac{d}{dx}\ln f_{q_s}(x). \tag{216}$$

Thus, from the definition of $r(x)$,

$$\frac{d}{dx}\lambda_\rho \mathcal{M}_{\lambda_\rho}[\,\rho^{\text{opt}}\,](x) = -q_s\frac{d}{dx}\ln(f_{q_s}(x)). \tag{217}$$

Integrating both sides of this equation, we find that

$$\mathcal{M}_{\lambda_\rho}[\,\rho^{\text{opt}}\,](x) = -\frac{q_s}{\lambda_\rho}\ln(f_{q_s}(x)) + b, \tag{218}$$

where b is an arbitrary constant. We now invert the Moreau envelope, using the property that for convex f, $\mathcal{M}_\lambda[\,f\,] = g$ implies $f = -\mathcal{M}_\lambda[\,-g\,]$ (derived in Appendix C.2) yields

$$\rho_{\text{opt}} = -\mathcal{M}_{\lambda_\rho}[\,\tfrac{q_s}{\lambda_\rho}\ln f_{q_s} - b\,] = -\mathcal{M}_{\lambda_\rho}[\,\tfrac{q_s}{\lambda_\rho}\ln f_{q_s}\,] + b. \tag{219}$$

Because additive constants will not effect on the M-estimation algorithm, we can write this as

$$\rho_{\text{opt}}(x) = -\mathcal{M}_{\lambda_\rho}[\,\tfrac{q_s}{\lambda_\rho}\ln f_{q_s}\,](x) = -\min_y\left[\frac{q_s}{\lambda_\rho}\ln f_{q_s}(y) + \frac{(x-y)^2}{2\lambda_\rho}\right] = -\frac{q_s}{\lambda_\rho}\mathcal{M}_{q_s}[\,\ln f_{q_s}\,](x). \tag{220}$$

The unregularized RS equations from Result 4 have an additional symmetry that we can exploit, the equations are unchanged under the transform $\rho \to c\rho$ and $\lambda_\rho \to \frac{1}{c}\lambda_\rho$ where c is a positive real number. Since we can re-scale the loss function without changing the optimal of the inference algorithm, we simplify the optimal loss function by choosing $\lambda_\rho = q_s$ so that

$$\rho_{\text{opt}} = -\mathcal{M}_{q_s}[\,\ln f_{q_s}\,]. \tag{221}$$

Result 17. *Given the loss function form:*

$$\rho = -\mathcal{M}_q[\,\ln P_{\epsilon_q}\,]. \tag{222}$$

The proximal map of ρ becomes:

$$\mathcal{P}_q[\,\rho\,](\epsilon_q) = \hat{\epsilon}^{MMSE}(\epsilon_q). \tag{223}$$

We first invert the Moreau envelope in (222) as explained in Appendix C.2, which yields

$$\mathcal{M}_q[\,\rho\,](x) = -\ln P_{\epsilon_q}. \tag{224}$$

We want to compute the proximal map of ρ, which is related to the Moreau envelope by the identity (367)

$$\mathcal{P}_q[\,\rho\,](x) = x - q\mathcal{M}_q[\,\rho\,]'(x). \tag{225}$$

To calculate the second term, we differentiate the form of the Moreau envelope and multiply by a constant:

$$-q\mathcal{M}_q[\,\rho\,]'(x) = q\frac{d}{dx}\ln P_{\epsilon_q}(x) = q\frac{\frac{d}{dx}\int\frac{1}{\sqrt{2\pi q_s}}e^{-\frac{(x-y)^2}{2q_s}}P_\epsilon(y)dy}{P_{\epsilon_q}(x)} = \frac{\int\frac{(y-x)}{\sqrt{2\pi q_s}}e^{-\frac{(x-y)^2}{2q_s}}P_\epsilon(y)dy}{P_{\epsilon_q}(x)} = \int(\epsilon-\epsilon_{q_s})P(\epsilon|\epsilon_{q_s})d\epsilon. \tag{226}$$

The final equality follows from Bayes rule: we can derive an expression for the prior on noise ϵ given a version corrupted by gaussian noise ϵ_q:

$$P(\epsilon|\epsilon_q) = \frac{P(\epsilon_q|\epsilon)P(\epsilon)}{P(\epsilon_q)} = \frac{1}{\sqrt{2\pi q}}\frac{e^{-\frac{(\epsilon_q-\epsilon)^2}{2q}}P(\epsilon)}{P(\epsilon_q)}. \tag{227}$$

It then follows that

$$\mathcal{P}_q[\,\rho\,](\epsilon_{q_s}) = \epsilon_{q_s} - q\mathcal{M}_q[\,\rho\,]'(\epsilon_{q_s}) = \int\epsilon P(\epsilon|\epsilon_{q_s}),d\epsilon = \hat{\epsilon}^{\text{MMSE}}(\epsilon_{q_s}), \tag{228}$$

which yields the desired result that the proximal map becomes an MMSE estimate.

Distribution of estimates

We let $q_s = q_s^{\text{opt}}$ be the optimal MSE for the remainder of this section for notational simplicity. We can use Result 17 to compute the distribution (67) of $\hat{\epsilon}$ in the unregularized optimal case where we can choose $\lambda_\rho = q_s$ so that,

$$\hat{\epsilon}^{\text{opt}}(\epsilon_{q_s}) = \mathcal{P}_{\lambda_\rho}[\,\rho^{\text{opt}}\,](\epsilon_{q_s}) = \mathcal{P}_{q_s}[\,\rho^{\text{opt}}\,](\epsilon_{q_s}) = \hat{\epsilon}^{\text{MMSE}}(\epsilon_{q_s}). \tag{229}$$

MSE of noise estimation

We define q_ϵ to be the per component MSE of an algorithm in estimating noise: $q_\epsilon = \left\langle\!\left\langle\,(\hat{\epsilon}-\epsilon)^2\,\right\rangle\!\right\rangle$. As we showed in (83), in the optimal unregularized case,

$$q_\epsilon = q_s\left(\frac{\alpha-1}{\alpha}\right). \tag{230}$$

From (215), in the optimal case α satisfies

$$\alpha = \frac{1}{q_s J\left[\,\epsilon_{q_s}\,\right]}. \tag{231}$$

Combining these yields

$$q_\epsilon = q_s - q_s^2 J\left[\,\epsilon_{q_s}\,\right] = q_\epsilon^{\text{MMSE}}(q_s). \tag{232}$$

The final equality follows from the relation between scalar Fisher information and MMSE given in Appendix B.4.

Training and test errors

Next we compute the training and test error. The test error is simply

$$\mathcal{E}^{\text{test}}_{\text{opt}} = q_s + \sigma_\epsilon^2 = \frac{1}{\alpha J\left[\,\epsilon_{q_s}\,\right]} + \sigma_\epsilon^2. \tag{233}$$

We can also compute the training error

$$\mathcal{E}^{\mathrm{train}}_{\mathrm{opt}} = \left\langle\!\left\langle \left(\hat{\epsilon}^{\mathrm{opt}}\right)^2 \right\rangle\!\right\rangle. \tag{234}$$

If we first compute the $q_\epsilon^{\mathrm{MMSE}}(q_s)$,

$$q_\epsilon^{\mathrm{MMSE}}(q_s) = \left\langle\!\left\langle \int (\epsilon - \hat{\epsilon}^{\mathrm{opt}}(\epsilon_{q_s}))^2 P(\epsilon|\epsilon_{q_s}) d\epsilon \right\rangle\!\right\rangle_{\epsilon_{q_s}} = \left\langle\!\left\langle \int \epsilon^2 P(\epsilon|\epsilon_{q_s}) \right\rangle\!\right\rangle - \left\langle\!\left\langle \left(\hat{\epsilon}^{\mathrm{opt}}(\epsilon_{q_s})\right)^2 \right\rangle\!\right\rangle_{\epsilon_{q_s}} = \sigma_\epsilon^2 - \mathcal{E}^{\mathrm{train}}_{\mathrm{opt}}, \tag{235}$$

thus

$$\mathcal{E}^{\mathrm{train}}_{\mathrm{opt}} = \sigma_\epsilon^2 - q_\epsilon^{\mathrm{MMSE}}(q_s). \tag{236}$$

5.2 Regularized inference

Result 18. *Under the assumption of log-concave probability densities P_ϵ and P_s, we find the optimal loss and regularizer pair to be:*

$$\rho^{opt} = -\mathcal{M}_{q_s}\left[\,\ln\left(P_{\epsilon_{q_s}}\right)\right] \qquad \sigma^{opt} = -\mathcal{M}_{q_d}\left[\,\ln\left(P_{s_{q_d}}\right)\right] \tag{237}$$

and q_s, q_d are chosen such that q_s is the minimum non-negative order parameter satisfying

$$q_d = \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s} z\right]}, \qquad q_s = q_d\left(1 - q_d J\left[s^0 + \sqrt{q_d} z\right]\right) = q_s^{MMSE}(q_d). \tag{238}$$

The measurement deviations and parameter estimates take on a value

$$\hat{\epsilon} = \hat{\epsilon}^{MMSE}(\epsilon_{q_s}) = \int \epsilon P(\epsilon|\epsilon_{q_s}) d\epsilon, \tag{239}$$

$$\hat{s} = \hat{s}^{MMSE}(s^0_{q_d}) = \int s P(s|s^0_{q_d}) ds, \tag{240}$$

with probability $P(s^0_{q_d})$, $P(\epsilon_{q_s})$, moreover the MSE of the optimal M-estimator in estimating ϵ is

$$q_\epsilon = \left\langle\!\left\langle (\hat{\epsilon}(\epsilon_{q_s}) - \epsilon)^2 \right\rangle\!\right\rangle = q_\epsilon^{MMSE}(q_s). \tag{241}$$

The corresponding test and training errors are:

$$\mathcal{E}^{test} = \sigma_\epsilon^2 + q_s = \sigma_\epsilon^2 + q_s^{MMSE}(q_d) \qquad \mathcal{E}^{train} = \sigma_\epsilon^2 - q_\epsilon = \sigma_\epsilon^2 - q_\epsilon^{MMSE}(q_s), \tag{242}$$

Derivation of the Algorithm
Our goal is to determine the optimal pair of functions ρ, σ, which we write as $\rho^{\mathrm{opt}}, \sigma^{\mathrm{opt}}$, to minimize the MSE of our estimation. We are optimizing simultaneously over regularizer and loss function to minimize q_s under the RS constraints (70 - 73). To simplify notation, we will let $g = P_s$ and noise $f = P_\epsilon$, and letting f_q and g_a be the probability densities of random variables $\epsilon + \sqrt{q} z_\epsilon$ and $s^0 + \sqrt{a} z_s$, respectively for independent unit gaussian random variables z_ϵ, z_s. Under the definition $r_\rho(x) = \lambda_\rho \mathcal{M}_{\lambda_\rho}[\,\rho\,]'(x)$ and $r_\sigma(x) = \lambda_\sigma \mathcal{M}_{\lambda_\sigma}[\,\sigma\,]'(x)$, we write these constraints as

Thus,

$$\frac{q_d}{\lambda_\sigma^2} - \frac{\alpha}{\lambda_\rho^2} \int f_{q_s}(y) r_\rho(y)^2 dy = 0 \tag{243}$$

$$\frac{1}{\lambda_\sigma} - \frac{\alpha}{\lambda_\rho} \int f_{q_s}(y) r_\rho'(y) dy = 0 \tag{244}$$

$$q_s - q_d - \int g_{q_d}(y)\left(r_\sigma(y)^2 - 2 q_d r_\sigma'(y)\right) dy = 0 \tag{245}$$

$$\lambda_\rho - \lambda_\sigma + \lambda_\sigma \int g_{q_d}(y) r_\sigma'(y) dy = 0 \tag{246}$$

we may define a Lagrangian for minimizing q_s under the constraints (243,244,245,246) as

$$\begin{aligned} L \quad &= q_s + \gamma_1\left[\frac{q_d}{\lambda_\sigma^2} - \frac{\alpha}{\lambda_\rho^2}\int f_{q_s}(y) r_\rho(y)^2 dy\right] + \gamma_2\left[\frac{1}{\lambda_\sigma} - \frac{\alpha}{\lambda_\rho}\int f_{q_s}(y) r_\rho'(y) dy\right] \\ &\gamma_3\left[q_s - q_d - \int g_{q_d}(y)\left(r_\sigma(y)^2 - 2q_d r_\sigma'(y)\right) dy\right] + \gamma_4\left[\lambda_\rho - \lambda_\sigma + \lambda_\sigma \int g_{q_d}(y) r_\sigma'(y) dy\right], \end{aligned} \tag{247}$$

where we have introduced lagrange parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$. To apply calculus of variation, we re-write the lagrangian as

$$L = q_s + \frac{\gamma_1 q_d}{\lambda_\sigma^2} + \frac{\gamma_2}{\lambda_\sigma} + \gamma_3 q_s - \gamma_3 q_d + \gamma_4 \lambda_\rho - \gamma_4 \lambda_\sigma + \int \left(S_\rho(y) + S_\sigma(y)\right) dy, \tag{248}$$

defining

$$S_\rho\left(y\right) = -\alpha\frac{f_{q_s}\left(y\right)}{\lambda_\rho}\left(\frac{\gamma_1}{\lambda_\rho}r_\rho\left(y\right)^2 + \gamma_2 r'_\rho\left(y\right)\right) \tag{249}$$

$$S_\sigma\left(y\right) = -g_{q_d}\left(y\right)\left(\gamma_3 r_\sigma\left(y\right)^2 - (\gamma_4\lambda_\sigma + 2\gamma_3 q_d)r'_\sigma\left(y\right)\right). \tag{250}$$

Requiring stationarity in both δr_ρ and δr_σ leads to the pair of E-L equations:

$$\frac{\partial S_\rho(r_\rho, r'_\rho; y)}{\partial r_\rho} = \frac{d}{dy}\frac{\partial S_\rho(r_\rho, r'_\rho; y)}{\partial r'_\rho}, \tag{251}$$

$$\frac{\partial S_\sigma(r_\sigma, r'_\sigma; y)}{\partial r_\sigma} = \frac{d}{dy}\frac{\partial S_\sigma(r_\sigma, r'_\sigma; y)}{\partial r'_\sigma}, \tag{252}$$

which lead to expressions r_ρ and r_σ:

$$r_\rho\left(y\right) = \lambda_\rho\frac{\gamma_2}{2\gamma_1}\frac{d}{dy}\ln\left(f_{q_s}\left(y\right)\right) \tag{253}$$

$$r_\sigma\left(y\right) = -\left(\frac{\gamma_4\lambda_\sigma}{2\gamma_3} + q_d\right)\frac{d}{dy}\ln\left(g_{q_d}\left(y\right)\right) \tag{254}$$

Substituting this form of r_ρ and r_σ into (249) and (250), and expressing the results in terms of the Fisher information, we find

$$\int S_\rho\left(y\right)dy = \alpha\frac{{\gamma_2}^2}{4\gamma_1}J\left[\epsilon + \sqrt{q_s}z\right] \tag{255}$$

$$\int S_\sigma\left(y\right)dy = \gamma_3\left(\frac{\gamma_4\lambda_\sigma}{2\gamma_3} + q_d\right)^2 J\left[s^0 + \sqrt{q_d}z\right]. \tag{256}$$

Thus, the Lagrangian (248) becomes

$$L = q_s + \frac{\gamma_1 q_d}{\lambda_\sigma^2} + \frac{\gamma_2}{\lambda_\sigma} + \gamma_3 q_s - \gamma_3 q_d + \gamma_4\lambda_\rho - \gamma_4\lambda_\sigma + \alpha\frac{{\gamma_2}^2}{4\gamma_1}J\left[\epsilon + \sqrt{q_s}z\right] + \gamma_3\left(\frac{\gamma_4\lambda_\sigma}{2\gamma_3} + q_d\right)^2 J\left[s^0 + \sqrt{q_d}z\right]. \tag{257}$$

The optimization problem may be further simplified by extremizing with respect to γ_1 and γ_2, which yields

$$\frac{q_d}{\lambda_\sigma^2} = \frac{\alpha{\gamma_2}^2}{4{\gamma_1}^2}J\left[\epsilon + \sqrt{q_s}z\right] \tag{258}$$

$$\frac{q_d}{\lambda_\sigma} = -\frac{\gamma_2}{2\gamma_1}J\left[\epsilon + \sqrt{q_s}z\right] \tag{259}$$

respectively. By combining these, we arrive at the constraints:

$$\frac{\gamma_2}{2\gamma_1} = -\frac{q_d}{\lambda_\sigma}. \tag{260}$$

$$q_d = \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s}z\right]} \tag{261}$$

If we next extremize with respect to λ_ρ, we find that $\gamma_4 = 0$. This suggests that constraint related to γ_4 does not impact the minimization problem. If we extremize with respect to γ_3 and γ_4 we find:

$$q_s - q_d + q_d^2 J\left[s^0 + \sqrt{q_d}z\right] = 0 \tag{262}$$

$$\lambda_\sigma - \lambda_\rho = q_d\lambda_\sigma J\left[s^0 + \sqrt{q_d}z\right] \tag{263}$$

Combining (262) and (263) yields an equation determining one order parameter given the other 3:

$$\frac{q_s}{q_d} = \frac{\lambda_\rho}{\lambda_\sigma}. \tag{264}$$

It follows that the optimization problem may be written in terms of only q_d, q_s as a minimum of q_s subject to constraints (261),(262). By defining

$$q_d(q_s) = \frac{1}{\alpha J\left[\epsilon + \sqrt{q_s}z\right]}, \tag{265}$$

the minimum MSE q_s^{opt} is the minimum value of q_s satisfying (262):

$$q_s = q_d - q_d^2 J\left[\, s^0 + \sqrt{q_d}z\,\right]. \tag{266}$$

For convenience, we define $q_d^{\text{opt}} = q_d(q_s^{\text{opt}})$. Note that the form of (266) implies that q_s is the MMSE of reconstructing s^0 from $s^0 + \sqrt{q_d}z$ where z is unit gaussian noise (see Appendix B.4 for a relationship between Fisher Information and MMSE).

To find the form of the optimal ρ, σ, we first combine (260) and (264) to reduce (253) and similarly (254) to show

$$\frac{d}{dy}\mathcal{M}_{\lambda_\rho}[\,\rho^{\text{opt}}\,](y) = -\frac{q_s^{\text{opt}}}{\lambda_\rho}\frac{d}{dy}\ln(f_{q_s^{\text{opt}}}(y)), \tag{267}$$

$$\frac{d}{dy}\mathcal{M}_{\lambda_\sigma}[\,\sigma^{\text{opt}}\,](y) = -\frac{q_d^{\text{opt}}}{\lambda_\sigma}\frac{d}{dy}\ln(g_{q_d^{\text{opt}}}(y)), \tag{268}$$

We can use the property derived in Appendix C.2 that $\mathcal{M}_\lambda[\,f\,] = g$ implies $f = -\mathcal{M}_\lambda[\,-g\,]$ under the assumption that f and g are convex along with the fact that the constants introduced by integrating both sides of the above equation will not effect the M-estimator, to write

$$\rho^{\text{opt}} = -\mathcal{M}_{\lambda_\rho}[\,\tfrac{q_s^{\text{opt}}}{\lambda_\rho}\ln(f_{q_s^{\text{opt}}})\,], \tag{269}$$

$$\sigma^{\text{opt}} = -\mathcal{M}_{\lambda_\sigma}[\,\tfrac{q_d^{\text{opt}}}{\lambda_\sigma}\ln(g_{q_d^{\text{opt}}})\,]. \tag{270}$$

Or equivalently,

$$\rho^{\text{opt}} = -\frac{q_s^{\text{opt}}}{\lambda_\rho}\mathcal{M}_{q_s^{\text{opt}}}[\,\ln(f_{q_s^{\text{opt}}})\,] \tag{271}$$

$$\sigma^{\text{opt}} = -\frac{q_d^{\text{opt}}}{\lambda_\sigma}\mathcal{M}_{q_d^{\text{opt}}}[\,\ln(g_{q_d^{\text{opt}}})\,]. \tag{272}$$

The regularized RS equations from Result 3 have an additional symmetry that we can exploit, the equations are unchanged under the transform $(\rho,\sigma) \to (c\rho, c\sigma)$ with $(\lambda_\rho, \lambda_\sigma) \to \left(\frac{1}{c}\lambda_\rho, \frac{1}{c}\lambda_\sigma\right)$ where c is a positive real number. Since we can multiply σ^{opt} and ρ^{opt} by the same positive constant without effecting the results of the inference algorithm, we simplify the optimal M-estimator by rescaling such that $\lambda_\rho = q_s^{\text{opt}}$, which from (264) implies $\lambda_\sigma = q_d^{\text{opt}}$, so that

$$\rho^{\text{opt}} = -\mathcal{M}_{q_s^{\text{opt}}}[\,\ln(f_{q_s^{\text{opt}}})\,], \tag{273}$$

$$\sigma^{\text{opt}} = -\mathcal{M}_{q_d^{\text{opt}}}[\,\ln(g_{q_d^{\text{opt}}})\,]. \tag{274}$$

Distribution of Estimates

For ease of notation let $q_s = q_s^{\text{opt}}$ and $q_d = q_d^{\text{opt}}$ for the remainder of this section. As we derived previously (63,64), the distribution measurement deviation $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\mathbf{s}}$ and of estimates $\hat{\mathbf{s}}$ are

$$\hat{\epsilon}^{\text{opt}}(\epsilon_{q_s}) = \mathcal{P}_{\lambda_\rho}[\,\rho^{\text{opt}}\,](\epsilon_{q_s}), \tag{275}$$

$$\hat{s}^{\text{opt}}(s^0_{q_d}) = \mathcal{P}_{\lambda_\sigma}[\,\sigma^{\text{opt}}\,](s^0_{q_d}). \tag{276}$$

Since we can set $\lambda_\rho = q_s$, $\lambda_\sigma = q_d$ in the optimal case, we can apply Result 17, and find that

$$\hat{\epsilon}^{\text{opt}}(\epsilon_{q_s}) = \mathcal{P}_{q_s}[\,\rho^{\text{opt}}\,](\epsilon_{q_s}) = \hat{\epsilon}^{\text{MMSE}}(\epsilon_{q_s}), \tag{277}$$

$$\hat{s}^{\text{opt}}(s^0_{q_d}) = \mathcal{P}_{\lambda_\sigma}[\,\sigma^{\text{opt}}\,](s^0_{q_d}) = \hat{s}^{\text{MMSE}}(s^0_{q_d}). \tag{278}$$

It also follows that

$$q_s^{\text{MMSE}}(q_d) = \left\langle\!\left\langle \int (s^0 - \hat{s}^{\text{opt}})^2 P(s^0|s^0_{q_d})ds^0 \right\rangle\!\right\rangle_{s^0_{q_d}} = \left\langle\!\left\langle \left(s^0 - \hat{s}^{\text{opt}}\right)^2 \right\rangle\!\right\rangle = q_s, \tag{279}$$

so q_s is the MMSE of inferring s^0 from $s^0_{q_d}$.

MSE of Noise Estimation

As we showed in (82), in the optimal regularized case,

$$q_\epsilon = q_s + \frac{q_s}{\alpha}\left[\frac{q_d}{q_s}\left(\frac{\lambda_\rho}{\lambda_\sigma} - \frac{q_s}{q_d}\right)^2 - \frac{q_s}{q_d}\right]. \tag{280}$$

Choosing $\lambda_\rho = q_s$, $\lambda_\sigma = q_d$ and substituting the form of q_d (265), this expression can be simplified as

$$q_\epsilon = q_s - q_s^2 J[\,\epsilon_{q_s}\,] = q_\epsilon^{\text{MMSE}}(q_s). \tag{281}$$

See the relation between scalar Fisher information and MMSE in Appendix B.4.

Training and Test Errors

The test error is simply

$$\mathcal{E}^{\text{test}} = \sigma_\epsilon^2 + q_s = \sigma_\epsilon^2 + q_s^{\text{MMSE}}(q_d). \tag{282}$$

From the definition of training error, we have that

$$\mathcal{E}^{\text{train}} = \left\langle\!\left\langle \hat{\epsilon}^2 \right\rangle\!\right\rangle_{\epsilon_{q_s}}, \qquad \mathcal{E}^{\text{train}}_{\text{opt}} = \left\langle\!\left\langle \left(\hat{\epsilon}^{\text{opt}}\right)^2 \right\rangle\!\right\rangle_{\epsilon_{q_s}}. \tag{283}$$

It then follows from (277) that

$$q_\epsilon^{\text{MMSE}}(q_s) = \left\langle\!\left\langle \int (\epsilon - \hat{\epsilon}^{\text{opt}}(\epsilon_{q_s}))^2 P(\epsilon|\epsilon_{q_s})d\epsilon \right\rangle\!\right\rangle_{\epsilon_{q_s}} = \left\langle\!\left\langle \int \epsilon^2 P(\epsilon|\epsilon_{q_s}) \right\rangle\!\right\rangle - \left\langle\!\left\langle \left(\hat{\epsilon}^{\text{opt}}(\epsilon_{q_s})\right)^2 \right\rangle\!\right\rangle_{\epsilon_{q_s}} = \sigma_\epsilon^2 - \left\langle\!\left\langle \left(\hat{\epsilon}^{\text{opt}}(\epsilon_{q_s})\right)^2 \right\rangle\!\right\rangle_{\epsilon_{q_s}}. \tag{284}$$

Rearranging, and substituting the form of $\mathcal{E}^{\text{train}}_{\text{opt}}$, it immediately follows that

$$\mathcal{E}^{\text{train}}_{\text{opt}} = \sigma_\epsilon^2 - q_\epsilon^{\text{MMSE}}(q_s). \tag{285}$$

5.3 Analytic bounds on MSE

Result 19. *The accuracy of unregularized M-estimation is invariant with respect to the true s^0 distribution and the optimal MSE obeys:*

$$\frac{1}{\sigma_\epsilon^2 J\left[\,\epsilon\,\right]}\frac{1}{\alpha-1} \leq \frac{q_s^{opt}}{\sigma_\epsilon^2} \leq \frac{1}{\alpha-1}. \tag{286}$$

In the regularized case, the optimal MSE $\bar{q}_s^{opt} = \frac{q_s^{opt}}{\sigma_s^2}$ satisfies the bounds

$$\frac{1}{\sigma_s^2 J\left[\,s^0\,\right]}\bar{q}_s^{Quad}\left(\alpha, \frac{J\left[\,s^0\,\right]}{J\left[\,\epsilon\,\right]}\right) \leq \bar{q}_s^{opt} \leq \bar{q}_s^{Quad}\left(\alpha, \frac{\sigma_\epsilon^2}{\sigma_s^2}\right), \tag{287}$$

where $\bar{q}_s^{Quad}\left(\,\alpha, \Phi\,\right)$, the form of which is derived in (118), *is the positive root of the quadratic*

$$x^2 + (\Phi + \alpha - 1)x - \Phi = 0. \tag{288}$$

Note that under the condition of Gaussian signal and noise $J\left[\,s^0\,\right] = \frac{1}{\sigma_s^2}, J\left[\,\epsilon\,\right] = \frac{1}{\sigma_\epsilon^2}$, thus

$$\bar{q}_s^{opt} = \bar{q}_s^{Quad}\left(\alpha, \frac{\sigma_\epsilon^2}{\sigma_s^2}\right). \tag{289}$$

Derivation.— In the unregularized case, quadratic M-estimation (least-squares regression) has the form (116) which must upper bounds the optimal MSE. The high dimensional Cramer-Rao bound (149) lower bounds the MSE.

The upper bound on the optimal regularized M-estimator follows from Result 9 for the optimal quadratic M-estimator, which must necessarily have an error greater than or equal to that of the optimal M-estimator. To derive our lower bound, we apply the convolutional Fisher inequality to the bounds derived for regularized M-estimation:

$$q_d \geq \frac{1}{\alpha J\left[\,\epsilon_{q_s}\,\right]} \geq \frac{1}{\alpha}\left(\frac{1}{J\left[\,\epsilon\,\right]} + q_s\right), \tag{290}$$

$$q_s \geq q_d\left(1 - q_d J\left[\,s^0_{q_d}\,\right]\right) \geq q_d\left(1 - \frac{q_d J\left[\,s^0\,\right]}{q_d J\left[\,s^0\,\right] + 1}\right). \tag{291}$$

Combining the previous two inequalities yields the desired lower bound.

5.4 Limiting cases

Result 20. ***Limiting Case for Optimal M-estimation***: *The form of the optimal M-estimator derived for log concave parameter and noise distributions further simplifies under various limits of the measurement density α and SNR:*

High SNR

Case $\alpha > 1$:

$$q_s \approx \frac{1}{\alpha J\left[\,\epsilon_{q_s}\,\right]},$$

Case $\alpha \ll 1$:

$$q_s \approx \sigma_s^2 \bar{q}_s^{Quad}\left(\alpha, \frac{\sigma_\epsilon^2}{\sigma_s^2}\right)$$

***SNR is* $\mathbf{O(1)}$**

Case $\alpha \gg 1$:

$$q_s \approx \frac{1}{\alpha J\left[\,\epsilon\,\right]}.$$

Case $\alpha \ll 1$:

$$q_s \approx \frac{1}{\frac{1}{\sigma_s^2} + \alpha J\left[\,\epsilon_{q_s}\,\right]}$$

Low SNR

Case of finite α:

$$q_s \approx \frac{1}{\frac{1}{\sigma_s^2} + \alpha J\left[\,\epsilon\,\right]}.$$

Derivation.— We will choose σ_s to be $O(1)$ and vary the scale of the noise σ_ϵ corresponding to different limits of SNR. We will also assume that P_ϵ, P_s have finite Fisher information when their variances are finite. From Result 18, in cases when the optimal M-estimator may be computed, MSE q_s obeys the conditions:

$$q_d = \frac{1}{\alpha J\left[\,\epsilon_{q_s}\,\right]}, \tag{292}$$

$$q_s = q_d(1 - q_d J\left[\,s^0_{q_d}\,\right]). \tag{293}$$

High SNR ($\sigma_\epsilon \to 0$)

Case $\alpha > 1$: In the over-sampled regime, where $\alpha > 1$, even least-squares regression will achieve reconstruction with an error proportional to σ_ϵ, as can be seen in (116). The optimal M-estimator performance has a strictly lower error than least-squares, thus $q_s \to 0$. It follows that $J\left[\, \epsilon_{q_s} \right]$ approaches infinity, so from (292) we know $q_d \to 0$. It follows from (293) that in this limit of small q_s and q_d, $q_s \approx q_d$. Substituting this into (292) yields

$$q_s \approx \frac{1}{\alpha J\left[\, \epsilon_{q_s} \right]}. \tag{294}$$

Of course, we need σ_ϵ sufficiently small so that $q_s \approx 0$, and it is sufficient to bound q_s from above with the error from least squares: $q_s^{\text{LS}} = \frac{\sigma_\epsilon^2}{\alpha-1}$, we thus require lower noise when α is close to one such that $\sigma_\epsilon^2 \ll \alpha - 1$. A regularizer is not required to accurately estimate q_s in this regime, so information about the parameter distribution is not useful and we find that the high dimensional unregularized bound is an accurate approximation of achievable MSE.

Case $\alpha \ll 1$: q_s will be bounded above by σ_s^2 since it must outperform the trivial estimator which simply estimates zero for every parameter. q_s will also be non-zero for finite $J\left[\, s^0 \right]$, as can be seen from the lower bound from (287). It follows that for $\alpha \to 0$,

$$q_d = \frac{1}{\alpha J\left[\, \epsilon_{q_s} \right]} \to \infty. \tag{295}$$

In this limit, $s^0_{q_d}$ may be approximated as a gaussian with variance $q_d + \sigma_s^2$. Similarly q_s is $O(1)$ and the variance of ϵ is small, so ϵ_{q_s} is also approximately gaussian and

$$J\left[\, \epsilon_{q_s} \right] \approx \frac{1}{q_s + \sigma_\epsilon^2}, \qquad J\left[\, s^0_{q_d} \right] \approx \frac{1}{q_d + \sigma_s^2}. \tag{296}$$

Substituting into the optimal M-estimator conditions and solving for q_s yields the result that the optimal M-estimator error is equivalent to that of the optimal quadratic loss-regularizer pair in this regime.

SNR is O(1) ($\sigma_\epsilon = O(1)$)

Case $\alpha \gg 1$: In the the highly over-sampled regime, again even least squares regression approaches zero error, so that the optimal M-estimator has zero error $q_s \to 0$. Since α is large

$$q_d \approx \frac{1}{\alpha J\left[\, \epsilon \right]} \to 0. \tag{297}$$

It follows from (293) that for small q_d, $q_s \approx q_d$:

$$q_s \approx \frac{1}{\alpha J\left[\, \epsilon_{q_s} \right]} \approx \frac{1}{\alpha J\left[\, \epsilon \right]}, \tag{298}$$

where the final approximation follows from the fact that q_s is small and σ_ϵ is order 1. Therefore this is a regime where the Cramer-Rao bound becomes tight and maximum likelihood becomes optimal.

Case $\alpha \ll 1$: It immediately follows from (292) that $q_d \to \infty$ which implies that $s^0_{q_d}$ can be approximated by a gaussian so (293) becomes

$$q_s \approx q_d \left(1 - \frac{q_d}{q_d + \sigma_s^2} \right) = \frac{1}{\alpha J\left[\, \epsilon_{q_s} \right] + \frac{1}{\sigma_s^2}}. \tag{299}$$

Low SNR ($\sigma_\epsilon \to \infty$)

Case $\alpha < \infty$: In this regime q_d approaches infinity, which follows from our assumption that for large σ_ϵ the Fisher information of the noise approaches zero. For large q_d, we may approximate $s^0_{q_d}$ as a gaussian and recover (299). Since $q_s = O(1)$, it is very small compared to ϵ, so that $\epsilon_{q_s} \approx \epsilon$:

$$q_s \approx \frac{1}{\alpha J\left[\, \epsilon \right] + \frac{1}{\sigma_s^2}}. \tag{300}$$

5.5 Optimal noise-free inference

Result 21. *In the case of noise-free reconstruction (for P_s log-concave) the optimal regularization function is*

$$\sigma^{opt}(x) = -\mathcal{M}_{q_d}[\,\ln P_{s^0_{q_d}}\,](x), \tag{301}$$

and yields a MSE of q_s, where q_s and q_d satisfy

$$q_s = q_s^{MMSE}(q_d), \tag{302}$$

$$q_d = \frac{q_s}{\alpha}, \tag{303}$$

and the estimated parameters $\hat{s}$ of the M-estimator are distributed such that they take the value

$$\hat{s} = \hat{s}^{MMSE}(s^0_{q_d}) \tag{304}$$

with probability $P(s^0_{q_d})$.

This result follows from minimizing the MSE q_s under the noise free error relations of Result 7. We have that $\lambda_\rho = \alpha\lambda_\sigma$, $q_s = \alpha q_d$, and by substituting this pair of identities into (245) and (246), we find

$$\alpha q_d - q_d - \int g_{q_d}(y)\left(r_\sigma(y)^2 - 2q_d r'_\sigma(y)\right)dy = 0, \tag{305}$$

$$1 + \frac{1}{\alpha}\left(\int g_{q_d}(y) r'_\sigma(y) dy - 1\right) = 0. \tag{306}$$

If we then follow the same procedure of constructing a Lagrangian and performing calculus of variations, as in the previous section, we arrive at the form of σ^{opt} above as well as the constraint on q_s:

$$q_s = \alpha \frac{1-\alpha}{J\left[\,s^0 + \sqrt{\frac{1}{\alpha} q_s} z\,\right]}. \tag{307}$$

In order to remove α from the above equation, we first re-arrange it to solve for α:

$$\alpha = 1 - q_d J[\,s^0_{q_d}\,]. \tag{308}$$

Substituting the form of α into the preceding equation allows us to write q_s as

$$q_s = q_d\left(1 - q_d J[\,s^0_{q_d}\,]\right) = q_s^{\mathrm{MMSE}}(q_d), \tag{309}$$

where the final equality follows from relation between Fisher information and MMSE in Appendix B.4. We know from the regularized RS equations that the parameter distribution is distributed as

$$\hat{s} = \mathcal{P}_{\lambda_\sigma}[\,\sigma\,](s^0_{q_d}). \tag{310}$$

Applying Result 17 to the form of σ^{opt}, and using the fact that we can set $\lambda_\sigma = q_d$ for optimal inference, the distribution of estimated parameters (304) immediately follows. Note that it also follows directly from the form of the estimated parameters that

$$q_s = \left\langle\!\left\langle\, (\hat{s}^{\mathrm{MMSE}}(s^0_{q_d}) - s^0)^2 \,\right\rangle\!\right\rangle = q_s^{\mathrm{MMSE}}(q_d). \tag{311}$$

5.6 The limits of small and large smoothing of Moreau envelopes

Result 22. *We derive the form of the optimal M-estimator in the limits of both small and large smoothing. This yields simplified forms for both optimal unregularized and regularized inference at high and low measurement densities. Letting f_q denote the density of a pdf f convolved with a zero mean gaussian of variance q, and defining μ_f, σ_f^2 to be the mean and variance of f respectively, we find*:*

$$\lim_{q\to 0} \mathcal{M}_q[\,\ln f_q\,](x) = \ln f(x) \tag{312}$$

$$\lim_{q\to\infty} \mathcal{M}_q[\,\ln f_q\,](x) = \frac{(x-\mu_f)^2}{2\sigma_f^2}. \tag{313}$$

**A sufficient condition for (312) is that $\left|\frac{\partial f}{f}\right|$ is bounded above. For details on the subgradient ∂ see e.g. [4]. A sufficient condition for (313) is finite moments/cumulants of f.*

$$\mathcal{M}_q[\ln f_q](x) = \min_y \left[\ln f_q(y) + \frac{(x-y)^2}{2q}\right] = \left[\ln f_q\left(y_q^*(x)\right) + \frac{\left(x - y_q^*(x)\right)^2}{2q}\right] \tag{314}$$

where $y_q^*(x)$ is the minimizer of the Moreau envelop (the prox function), and thus satisfies

$$y_q^*(x) = x + q\frac{\partial f_q\left(y_q^*(x)\right)}{f_q\left(y_q^*(x)\right)}. \tag{315}$$

In the limit $q \to 0$,

$$\lim_{q\to 0}\mathcal{M}_q[\ln f_q](x) = \lim_{q\to 0}\left[\ln f_q\left(x + q\frac{\partial f_q\left(y_q^*(x)\right)}{f_q\left(y_q^*(x)\right)}\right) + q\frac{\left(\frac{\partial f_q\left(y_q^*(x)\right)}{f_q\left(y_q^*(x)\right)}\right)^2}{2}\right] = \ln f_0(x) = \ln f(x), \tag{316}$$

where the second equality follows from the assumption of bounded $\left|\frac{\partial f}{f}\right|$ (so that $\lim_{q\to 0} q\left|\frac{\partial f}{f}\right| = 0$). For the case where $q \to \infty$,

$$2\pi f_q(y) = \int dk e^{iky} e^{-\frac{qk^2}{2}}\tilde{f}(k) = \int dk e^{iky - \frac{qk^2}{2} + \ln\tilde{f}(k)} = \int dk e^{-q\left(\frac{k^2}{2} - i\frac{ky}{q} - \frac{1}{q}\ln\tilde{f}(k)\right)}. \tag{317}$$

Thus, letting $H(k) = \ln\tilde{f}(k)$,

$$f_\infty(y) = \frac{1}{2\pi}\lim_{q\to\infty}\int dk e^{-q\left(\frac{k^2}{2} - i\frac{ky}{q} - \frac{1}{q}H(k)\right)}. \tag{318}$$

Now, assuming f has finite cumulants, we can calculate the above function via a saddle point argument that

$$f_\infty(y) = \lim_{q\to\infty} e^{-q\left(\frac{\hat{k}^2}{2} - i\frac{\hat{k}y}{q} - \frac{1}{q}H(\hat{k})\right)}\sqrt{\frac{1}{2\pi\left(q - H''(\hat{k})\right)}}, \tag{319}$$

where $\hat{k}$ is the maximizer of the integrand. A perturbative expansion of $\hat{k}$ gives

$$\hat{k} = \frac{iy + H'(0)}{q} + \frac{H''(0)\left(iy + H'(0)\right)}{q^2} + \cdots \tag{320}$$

It is not hard to show that $H(0) = 0$, $H'(0) = -i\mu$, and $H''(0) = -\sigma^2$, where μ and σ are respectively the mean and variance of f and find

$$f_\infty(y) = \lim_{q\to\infty}\frac{1}{\sqrt{2\pi(q+\sigma^2)}}e^{-\frac{(y-\mu)^2}{2(q+\sigma^2)}}. \tag{321}$$

Thus, the smoothed distribution tends to a gaussian, where we can approximate the original pdf f by a gaussian with identical variance and mean. We can now use this approximation of ζ_q to compute ρ in the large q limit

$$\rho_\infty(x) = \lim_{q\to\infty} -\inf_y\left[-\frac{(y-\mu)^2}{2(q+\sigma^2)} + \frac{(x-y)^2}{2q}\right]. \tag{322}$$

After some manipulation we find

$$\rho_\infty(x) = \frac{(x-\mu)^2}{2\sigma^2}. \tag{323}$$

We use this result in the main paper to demonstrate that under suitable assumptions in the high dimensional regime the optimal unregularized loss approaches a quadratic, and the optimal regularized M-estimator has a regularization function approaching a square penalty.

These results can be applied to simplify optimal inference in various limits. For example in unregularized inference, as $\alpha \to 1$, $q_s \to \infty$, so the results of this section imply that the optimal loss function (196) simplifies to a quadratic in the under-sampled limit. Also, as $\alpha \to \infty$, $q_s \to 0$, implying that the optimal loss function simplifies to maximum-likelihood in the over-sampled limit.

In the case of regularized inference, as $\alpha \to 0$, $q_d \to \infty$ so that the optimal regularization function (237) approaches a quadratic. If we are additionally in the low noise (high SNR) limit, then q_s is $O(1)$ while ϵ is small so that smoothing dominates and the optimal loss also approaches a quadratic. In the over-sampled limit of $\alpha \to \infty$, $q_d \to 0$ and $q_s = q_d - q_d^2 J\left[s^0_{q_d}\right]) \approx q_d$, since the smoothing for both the loss and regular approach 0 at the same rate the optimal M-estimator (237) approaches MAP.

A Review of Single Parameter Inference

A.1 Single parameter M-estimation

Consider the problem of inferring a single parameter s^0 given measurements y_μ for $\mu = 1, ..., N$:

$$y_\mu = s^0 + \epsilon_\mu, \tag{324}$$

where the noise ϵ_μ is drawn iid from a probability density function P_ϵ. We apply an unregularized M-estimator to estimate s^0:

$$\hat{s} = \arg\min \sum_\mu \rho(y_\mu - s) = \arg\min \sum_\mu \rho(\epsilon_\mu - (s - s^0)). \tag{325}$$

Thus $\hat{s}$ is the solution to,

$$\sum_\mu \rho'(\epsilon_\mu - (\hat{s} - s^0)) = 0. \tag{326}$$

For an unbiased M-estimator, in the large N limit $\hat{s} - s^0$ becomes small so that we can Taylor expand in this small quantity:

$$\sum_\mu \rho'(\epsilon_\mu) - \sum_\mu \rho''(\epsilon_\mu)(\hat{s} - s^0) = 0. \tag{327}$$

Rearranging and squaring both sides,

$$\left(\sum_\mu \rho''(\epsilon_\mu)\right)^2 (\hat{s} - s^0)^2 = \left(\sum_\mu \rho'(\epsilon_\mu)\right)^2 \to \sum_\mu \rho'(\epsilon_\mu)^2, \tag{328}$$

Here we neglect the off diagonal terms in the final square, as the terms in the sum are uncorrelated random variables, so the average of the off diagonal terms will vanish. Thus, in the asymptotic large N limit, after averaging over the noise, we find

$$q_s = (\hat{s} - s^0)^2 = \frac{\sum_\mu \rho'(\epsilon_\mu)^2}{\left(\sum_\mu \rho''(\epsilon_\mu)\right)^2} \to \frac{\langle\langle \rho'(\epsilon)^2 \rangle\rangle}{\langle\langle \rho''(\epsilon) \rangle\rangle^2}, \tag{329}$$

where we have define q_s to be the MSE of signal estimate.

A.2 Bounds on single parameter M-estimation

We can use this result on the form (329) of the MSE of an M-estimator to derive a bound on the MSE of M-estimation. We begin with the expression in the denominator and apply integration by parts,

$$\left(\int \rho''(x) P_\epsilon(x) dx\right)^2 = \left(\int \rho'(x) P'_\epsilon(x) dx\right)^2 = \left(\int \left(\rho'(x)\sqrt{P_\epsilon(x)}\right)\left(\frac{P'_\epsilon(x)}{\sqrt{P_\epsilon(x)}}\right) dx\right)^2 \leq \int P_\epsilon(x)\rho'(x)^2 dx \int \frac{P'_\epsilon(x)^2}{P_\epsilon(x)} dx, \tag{330}$$

where the inequality follows from Cauchy-Schwartz. The final integral is simply the Fisher information from (348). This bound will turn out to be a special case of a more general bound known as the Cramer-Rao bound, that applies any unbiased inference problem. Thus, our analytic asymptotic expression (329) for MSE is bounded as

$$q_s = \frac{\langle\langle \rho'(\epsilon)^2 \rangle\rangle}{\langle\langle \rho''(\epsilon) \rangle\rangle^2} \geq \frac{1}{\int \frac{P'_\epsilon(x)^2}{P_\epsilon(x)} dx} = \frac{1}{J[\epsilon]}. \tag{331}$$

The Cauchy-Schwartz inequality is saturated when the functions in the inner product (330) are proportional so that

$$\rho'(x)\sqrt{P_\epsilon(x)} \propto \frac{P'_\epsilon(x)}{\sqrt{P_\epsilon(x)}}. \tag{332}$$

It follows that $\rho(x) = -\log P_\epsilon(x)$ is the asymptotically optimal M-estimator and achieves the optimal minimum MSE amongst all unbiased estimators.

A.3 Asymptotic analysis of single parameter Bayesian MMSE estimation

Suppose we are given N noisy measurements indexed by $\mu = 1, ..., N$ of an unknown signal s^0:

$$y_\mu = s^0 + \epsilon_\mu. \tag{333}$$

We would like to estimate s^0 from the measurements y_μ. To this end we apply Bayes rule to compute a posterior distribution

$$P(s|\mathbf{y}) \propto \prod_{\mu=1}^{N} P_\epsilon(y_\mu - s) P_s(s) = e^{\sum_{\mu=1}^{N} \log(P_\epsilon(y_\mu - s))} P_s(s) = e^{\sum_{\mu=1}^{N} \log\left(P_\epsilon(\epsilon_\mu - (s - s^0))\right)} P_s(s). \tag{334}$$

For large N, the posterior distribution will be concentrated in the vicinity of s^0. Therefore we can expand the distribution around s^0, taking $s - s^0$ to be small:

$$\sum_{\mu=1}^{N} \log\left(P_\epsilon(\epsilon_\mu - (s - s^0))\right) \approx \sum_{\mu=1}^{N} \log'\left(P_\epsilon(\epsilon_\mu)\right)(s - s^0) + \frac{1}{2}\sum_{\mu=1}^{N} \log''\left(P_\epsilon(\epsilon_\mu)\right)(s - s^0)^2 \tag{335}$$

These Taylor series coefficients are random variables and as $N \to \infty$ the central limit theorem implies:

$$\sum_{\mu=1}^{N} \log'\left(P_\epsilon(\epsilon_\mu)\right) \to \mathcal{N}(0, NJ\left[\,\epsilon\,\right]), \tag{336}$$

$$\sum_{\mu=1}^{N} \log''\left(P_\epsilon(\epsilon_\mu)\right) \to NJ\left[\,\epsilon\,\right]. \tag{337}$$

Thus, in the large N limit, the precise measurement vector $\mathbf{y}$ in the prior distribution (334) may be replaced with random variables so that

$$P(s|s^0, z) \propto P_s(s) e^{-\frac{(s - s^0 - z\sqrt{q_d})^2}{2q_d}}, \tag{338}$$

where z is a unit gaussian random variable and we define $q_d = \frac{1}{NJ\left[\,\epsilon\,\right]}$, so that q_d is the unbiased error from the Classical Cramer-Rao bound, achieved by maximum likelihood in the large N limit.

$$\hat{s}^{\text{MMSE}} = \int s p_s(s|s^0, z) ds. \tag{339}$$

It also follows that the MMSE achievable given the data is simply

$$q_s = \left\langle\!\left\langle \int (s - \hat{s}^{\text{MMSE}})^2 p(s|s^0, z) ds \right\rangle\!\right\rangle_{s^0, z}. \tag{340}$$

Using the relationship between MMSE and Fisher information in Appendix B.4, the MMSE of inferring a random variable s^0 contaminated by gaussian noise of variance q_d, may be written in terms of the Fisher information:

$$q_s = q_s^{\text{MMSE}}(q_d) = q_d - q_d^2 J\left[\, s^0 + \sqrt{q_d} z \,\right]. \tag{341}$$

Then, the convolution Fisher inequality (157) implies

$$\frac{1}{J\left[\, s^0 + \sqrt{q_d} z \,\right]} \geq \frac{1}{J\left[\, s^0 \,\right]} + q_d, \tag{342}$$

so that

$$q_s = q_d - q_d^2 J\left[\, s^0 + \sqrt{q_d} z \,\right] \geq q_d \left(1 - \frac{q_d}{\frac{1}{J\left[\, s^0 \,\right]} + q_d}\right) = q_d \left(\frac{1}{1 + q_d J\left[\, s^0 \,\right]}\right) = \frac{1}{\frac{1}{q_d} + J\left[\, s^0 \,\right]} = \frac{1}{NJ\left[\,\epsilon\,\right] + J\left[\, s^0 \,\right]}, \tag{343}$$

yielding a Fisher information bound on asymptotic MMSE inference for a single parameter:

$$q_s \geq \frac{1}{NJ\left[\,\epsilon\,\right] + J\left[\, s^0 \,\right]}. \tag{344}$$

Special case of gaussian noise and parameters:

In the case of gaussian noise and parameters ($\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), s^0 \sim \mathcal{N}(0, \sigma_s^2)$), (344) holds with equality for all N (not just in the asymptotic limit):

$$q_s = \frac{1}{NJ\left[\,\epsilon\,\right] + J\left[\,s^0\,\right]}. \tag{345}$$

Note that $J\left[\,\epsilon\,\right] = \frac{1}{\sigma_\epsilon^2}$, $J\left[\,s^0\,\right] = \frac{1}{\sigma_s^2}$, and that we need not consider derivatives higher than second order in $\log(P_\epsilon(x))$, so we need only compute two terms: (336) and (337). In both cases the equalities hold exactly, since a sum of gaussian random variables is a gaussian random variable. Since we do not need to invoke the central limit theorem, this result holds for any N. Further, the convolutional Fisher inequality (342) becomes an equality for gaussian, this follows from that fact that the Fisher information is the inverse variance for a a gaussian:

$$\frac{1}{J\left[\,s^0 + \sqrt{q_d}z\,\right]} = \frac{1}{J\left[\,\sqrt{q_d + \sigma_s^2}z\,\right]} = q_d + \sigma_s^2 = q_d + \frac{1}{J\left[\,s^0\,\right]}. \tag{346}$$

B Fisher Information

B.1 Fisher information for scalar inference

The scalar Fisher information $J\left[\,\epsilon\,\right]$ quantifies the accuracy of any unbiased estimator to reconstruct a signal corrupted by additive noise ϵ drawn from probability density function P_ϵ. The problem is to estimate the scalar s^0 given a corrupted measurements y:

$$y = s^0 + \epsilon. \tag{347}$$

The definition of Fisher information in this scalar case is:

$$J\left[\,\epsilon\,\right] = \left\langle\!\left\langle \left(\tfrac{d}{ds^0}\ln p(y|s^0)\right)^2 \right\rangle\!\right\rangle_\epsilon = \int \frac{P_\epsilon'(z)^2}{P_\epsilon(z)}dz. \tag{348}$$

B.2 Convolutional Fisher inequality

The scalar Fisher information obeys the following inequality:

$$J\left[\,\epsilon_1 + \epsilon_2\,\right] \le \frac{J\left[\,\epsilon_1\,\right]J\left[\,\epsilon_2\,\right]}{J\left[\,\epsilon_1\,\right] + J\left[\,\epsilon_2\,\right]}, \tag{349}$$

where ϵ_1, ϵ_2 are drawn from separate probability densities. See [5] p.674.

B.3 Fisher information is minimized by a gaussian distribution

For a noise distribution P_ϵ, the Fisher information is bounded below as

$$J\left[\,\epsilon\,\right] \ge \frac{1}{\sigma_\epsilon^2}, \tag{350}$$

and for a fixed variance σ_ϵ^2 is minimized for a gaussian distribution. This result follows from the inequality:

$$1 = \left(\int P_\epsilon(x)\left(\frac{(x - \langle\,\epsilon\,\rangle)P_\epsilon'(x)}{P_\epsilon(x)}\right)dx\right)^2 \le \int (x - \langle\,\epsilon\,\rangle)^2 P_\epsilon(x)dx \int \frac{P_\epsilon'(x)^2}{P_\epsilon(x)}dx = \sigma_\epsilon^2 J\left[\,\epsilon\,\right], \tag{351}$$

where the left equality follows from integration by parts, the inequality follows from Cauchy Schwartz, and the Fisher information appears in the final equality by substituting the definition of Fisher Information (348). This proves the desired bound. Moreover, the cauchy-schwartz that the bound is saturated when

$$\frac{P_\epsilon'(x)}{P_\epsilon(x)} \propto \pm(x - \langle\,\epsilon\,\rangle). \tag{352}$$

Integrating both sides and choosing appropriate constants s.t. P_ϵ in a probability distribution yields

$$P_\epsilon(x) \propto e^{-\frac{(x-\langle\epsilon\rangle)^2}{2\sigma_\epsilon^2}}, \tag{353}$$

so that this bound is saturated only for a gaussian P_ϵ.

B.4 Relation between MMSE and Fisher information

The MMSE of estimating a single random variable s^0 from measurements corrupted by gaussian noise of variance q_d can be written in terms of the Fisher Information as

$$q_s^{\text{MMSE}}(q_d) = q_d - q_d^2 J\left[\, s^0 + \sqrt{q_d} z \,\right]. \tag{354}$$

See e.g. [6] equation 28, p. 38.

B.5 Cramer-Rao bound

For the case of multiple parameters, the Cramer-Rao bound is written in terms of the Fisher information matrix, we will state this bound here and derive the form of the Fisher information matrix for the problem considered in this work.

For any estimator $\mathbf{\hat{s}}$ of the true parameters $\mathbf{s^0}$ given the data $\mathbf{z}$, where the conditional probability of data given parameters is $p(\mathbf{z}|\mathbf{s^0})$, we define the Fisher information matrix as

$$I_{ij} = \left\langle\!\!\left\langle \tfrac{\partial}{\partial s_i^0} \log p(\mathbf{z}|\mathbf{s^0}) \tfrac{\partial}{\partial s_j^0} \log p(\mathbf{z}|\mathbf{s^0}) \right\rangle\!\!\right\rangle_z . \tag{355}$$

The Cramer-Rao bound states that for any unbiased inference algorithm with estimates $\mathbf{\hat{s}}$, so that

$$\left\langle\!\left\langle\, (\hat{s}_k - s_k^0)^2 \,\right\rangle\!\right\rangle_{\mathbf{z}} \geq I_{kk}^{-1}. \tag{356}$$

See [5] for a derivation of the Cramer-Rao bound.

Applying Cramer-Rao to the problem considered in this paper:

We first derive the form of the Fisher information matrix $\mathbf{I}$ for the system studied in this paper:

$$\mathbf{y} = \mathbf{X}\mathbf{s^0} + \boldsymbol{\epsilon}, \tag{357}$$

where $\epsilon_\mu \sim f$ iid, $\mathbf{y} \in \mathbb{R}^N$,$\mathbf{s^0} \in \mathbb{R}^P$, and $\mathbf{X} \in \mathbb{R}^{N\times P}$. We apply the definition of the Fisher information matrix (355) to derive its form:

$$I_{jk} = \left\langle\!\!\left\langle \tfrac{\partial}{\partial s_j} \ln p(\mathbf{X},\mathbf{y}|\mathbf{s}) \tfrac{\partial}{\partial s_k} \ln p(\mathbf{X},\mathbf{y}|\mathbf{s}) \right\rangle\!\!\right\rangle_{\mathbf{X},\mathbf{y}}. \tag{358}$$

In this problem,

$$\ln p(\mathbf{X},\mathbf{y}|\mathbf{s}) = \ln p(y|X,s) + \ln p(X) = \sum_{\mu=1}^{N} \ln f(y_\mu - \mathbf{X}_\mu \cdot \mathbf{s}) + \ln p(X), \tag{359}$$

where $\mathbf{X}_\mu$ is the μ^{th} row of the matrix $\mathbf{X}$. Differentiating (358) yields

$$I_{jk} = \left\langle\!\!\left\langle \textstyle\sum_{\mu=1}^{N} X_{\mu j} X_{\mu k} \left(\tfrac{f'(\epsilon_\mu)}{f(\epsilon_\mu)}\right)^2 \right\rangle\!\!\right\rangle_{\mathbf{X},\boldsymbol{\epsilon}}. \tag{360}$$

For $j \neq k$ the input matrices will be uncorrelated so $I_{jk} = 0$. For $j = k$ then,

$$I_{jj} = \alpha \int dx \frac{f'(x)^2}{f(x)}. \tag{361}$$

Thus, the Fisher information matrix has no dependence on $\mathbf{s^0}$ and has the form:

$$I_{ij} = \delta_{ij}\alpha \int \frac{(f'(z))^2}{f(z)} dz = \delta_{ij}\alpha J\left[\,\epsilon\,\right]. \tag{362}$$

Substituting the Fisher information matrix into the Cramer-Rao bound, yields the variance per parameter:

$$q_s \geq \frac{1}{\alpha J\left[\,\epsilon\,\right]}. \tag{363}$$

C Moreau envelope and proximal map

C.1 Relation between proximal map and Moreau envelope

The Moreau envelope is a functional mapping parameterized by a regularization scalar λ. It maps a function f to

$$\mathcal{M}_\lambda[\,f\,](x) = \min_y \left[\frac{(x-y)^2}{2\lambda} + f(y)\right]. \tag{364}$$

We will denote the special case of $\mathcal{M}_1[f]$ by $\mathcal{M}\,[\,f\,]$. Some properties that follows from this definition are that the minimizers of f and $\mathcal{M}_\lambda[\,f\,]$ are the same, and that the Moreau envelope is also lower bound on the function f. The closely related proximal map is defined as

$$\mathcal{P}_\lambda[\,f\,](x) = \arg\min_y \left[\frac{(x-y)^2}{2\lambda} + f(y)\right]. \tag{365}$$

The proximal map can be viewed as a gradient descent step along the Moreau envelope:

$$\mathcal{P}_\lambda[\,f\,](x) - x = -\lambda\mathcal{M}_\lambda[\,f\,]'(x). \tag{366}$$

To derive (366), we begin with the derivative of the Moreau envelope and show

$$\mathcal{M}_\lambda[\,f\,]'(x) = \frac{d}{dx}\min_y \left[\frac{(x-y)^2}{2\lambda} + f(y)\right] = \frac{d}{dx}\left[\frac{(x-\hat{y})^2}{2\lambda} + f(\hat{y})\right], \tag{367}$$

where $\hat{y}$ is the minimizer of the RHS argument of (367). Differentiating with respect to $\hat{y}$ yields 0 at the minimum, so the $\hat{y}$ may be effectively treated as a constant and we need only differentiate with respect to x, which yields

$$\mathcal{M}_\lambda[\,f\,]'(x) = \frac{x-\hat{y}}{\lambda}. \tag{368}$$

It follows that

$$\mathcal{M}_\lambda[\,f\,]'(x) = \frac{x-\mathcal{P}_\lambda[\,f\,](x)}{\lambda}. \tag{369}$$

C.2 Inverse of the Moreau envelope

Result 23. ***Procedure for inverting a Moreau envelope:*** *For $\lambda > 0$ and f a convex, lower semi-continuous function such that $\mathcal{M}_\lambda[\,f\,] = g$, the Moreau envelope can be inverted so that $f = -\mathcal{M}_\lambda[\,-g\,]$.*

To derive this result, we first consider the case of $\lambda = 1$, from which the $\lambda > 0$ case will follow. Our assumption that $\mathcal{M}_\lambda[\,f\,] = g$ implies

$$g(x) = \mathcal{M}\,[\,f\,](x) = \min_y \left[\frac{(x-y)^2}{2} + f(y)\right] = \frac{x^2}{2} + \min_y\left[-xy + \frac{y^2}{2} + f(y)\right] = \frac{x^2}{2} - \max_y\left[xy - \frac{y^2}{2} - f(y)\right]. \tag{370}$$

We now define the Fenchel conjugate $.^*$, which operates on a function h to yield $h^*(x) = \max_y\,[xy - h(y)]$. We then define, for notational simplicity, the function $p_2(x) = \frac{x^2}{2}$. With this notation, (370) reduces to

$$g = p_2 - (f + p_2)^* \tag{371}$$

The Fenchel-Moreau theorem [7] states that if h is a convex and lower semi-continuous function, then $h = (h^*)^*$. These properties are assumed true for f and will also hold for $f + p_2$ so that (371) may be inverted, yielding:

$$f = (p_2 - g)^* - p_2. \tag{372}$$

We now write f in terms of a Moreau envelope by expanding the previous expression:

$$f(x) = -\frac{x^2}{2} + \max_y\,[xy - \frac{y^2}{2} + g(y)] = -\min_y\left[\frac{(x-y)^2}{2} - g(y)\right] = -\mathcal{M}\,[\,-g\,](x). \tag{373}$$

Thus, $\mathcal{M}\,[\,f\,] = g$ implies $f = -\mathcal{M}\,[\,-g\,]$. To extend this to $\lambda \neq 1$, we use the identity

$$\lambda\mathcal{M}_\lambda[\,\tfrac{1}{\lambda}f\,] = \mathcal{M}\,[\,f\,], \tag{374}$$

which can be verified by substitution into the definition of the Moreau envelope (364). Combining the result $\mathcal{M}[\,f\,] = g$ implies $f = -\mathcal{M}[\,-g\,]$ with (374) yields that:

$$\mathcal{M}_\lambda[\,\tfrac{1}{\lambda}f\,] = \frac{1}{\lambda}g, \tag{375}$$

also implies

$$\frac{1}{\lambda}f = -\mathcal{M}_\lambda[\,-\tfrac{1}{\lambda}g\,], \tag{376}$$

which completes the derivation since $\frac{1}{\lambda}$ may be absorbed into the definition of f and g.

References

[1] Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014, 2013.

[2] D. Bean, PJ Bickel, N. El Karoui, and B. Yu. Optimal M-estimation in high-dimensional regression. *PNAS*, 110(36):14563–8, 2013.

[3] S. Ganguli and H. Sompolinsky. Statistical mechanics of compressed sensing. *Physical Review Letters*, 104(18):188701, May 2010.

[4] S Boyd and L Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[5] TM Cover and JA Thomas. *Elements of information theory.* Wiley and Sons, 2012.

[6] O Rioul. Information theoretic proofs of entropy power inequalities. *Information Theory, IEEE Transactions on*, 57(1):33–55, 2011.

[7] Shozo Koshi and Naoto Komuro. A generalization of the Fenchel-Moreau theorem. *Proceedings of the Japan Academy, Series A, Mathematical Sciences*, 59(5):178–181, 1983.