

10 Robust Supervised Learning and Optimization

Now we want to lift our understanding from mean estimation to general stochastic optimization problems of the form:

$$\min_{\theta \in \Theta} \ell(\theta; \mathcal{D}), \quad \text{where} \quad \ell(\theta; \mathcal{D}) = \mathbb{E}_{X \sim \mathcal{D}} [\ell(\theta; X)].$$

Here recall that $\Theta \subseteq \mathbb{R}^d$ is the parameter space, $\mathcal{D}$ is an unknown distribution over a data domain $\mathcal{X}$, and $\ell(\theta; x)$ measures the loss of parameter θ on a data point x . Given a sample $X = (X_1, \dots, X_n) \sim \mathcal{D}^n$, we want an algorithm $\mathcal{A}$ that outputs $\hat{\theta} = \mathcal{A}(X) \in \Theta$ with small excess risk

$$\ell(\hat{\theta}; \mathcal{D}) - \ell(\theta^*; \mathcal{D}), \quad \theta^* \in \arg \min_{\theta \in \Theta} \ell(\theta; \mathcal{D}).$$

The way we solve these kinds of stochastic optimization problems in practice (at least when training models) is empirical risk minimization (ERM), where we try to minimize the empirical loss

$$\ell(\hat{\theta}; X) - \ell(\theta^*; X) \quad \theta^* \in \arg \min_{\theta \in \Theta} \ell(\theta; X).$$

There are lots of ways to optimize functions, but in the context of machine learning, some form of gradient descent is the overwhelmingly most popular choice. At a high level, gradient descent is an iterative algorithm that starts with some initial guess θ_0 and repeatedly updates the current guess θ_t in the direction of the negative gradient of the loss function, which points in the direction of steepest descent. The update rule can be written as:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \ell(\theta_t; X),$$

where η is a step size parameter that controls how far we move in the direction of the negative gradient at each iteration. The process continues until we reach some kind of convergence criterion, or just get bored and give up.

The key thing that makes it possible to lift our robust algorithms from mean estimation to general optimization is that the gradient is a linear operator,

$$\nabla_{\theta} \ell(\theta; X) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \ell(\theta; X_i), \quad \nabla_{\theta} \ell(\theta; \mathcal{D}) = \mathbb{E}_{\mathcal{D}} [\nabla_{\theta} \ell(\theta; X)].$$

So if we just replace each point x with its gradient at a point θ , we can apply robust mean estimation to the resulting distribution. Note that this distribution may not be Gaussian, but we still get *some* mean estimation problem to solve.

Thus, the basic approach we will take is to run gradient descent on the empirical loss, but using a robust mean estimator to compute the gradient at each step. This gives us a robust version of gradient descent that we can try to prove things about. Gradient descent itself notoriously can be analyzed in a ton of different ways under a ton of different assumptions. But we can start with a basic setting just to give an example of how the pieces fit together.

10.1 Gradient Descent for Convex ERM

Fix a sample X and write $f(\theta) = \ell(\theta; X)$ for its empirical loss. A domain Θ is *convex* if, for every $\theta, \theta' \in \Theta$ and $\lambda \in [0, 1]$, the point $(1 - \lambda)\theta + \lambda\theta'$ also belongs to Θ . A function f is *convex* on Θ if

$$f((1 - \lambda)\theta + \lambda\theta') \leq (1 - \lambda)f(\theta) + \lambda f(\theta').$$

For a differentiable convex function, this is equivalent to the first-order inequality

$$f(\theta') \geq f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle.$$

In words, the tangent plane at any point lies below the function everywhere on the domain.

We assume Θ is nonempty, closed, and convex, and has *radius at most* R , meaning

$$\|\theta\|_2 \leq R \quad \text{for every } \theta \in \Theta.$$

Closedness ensures that the Euclidean projection onto Θ is well defined. We also assume f is convex and differentiable, with gradients bounded by L :

$$\|\nabla f(\theta)\|_2 \leq L \quad \text{for every } \theta \in \Theta.$$

This makes f *L-Lipschitz* on Θ : $|f(\theta) - f(\theta')| \leq L\|\theta - \theta'\|_2$. These properties hold, for example, if every pointwise loss $\ell(\cdot; X_i)$ is convex and differentiable with gradient norm at most L .

An ordinary gradient step can leave the domain, so we project each update back onto Θ . Write

$$\Pi_{\Theta}(y) = \arg \min_{\theta \in \Theta} \|\theta - y\|_2^2$$

for the Euclidean projection.

Algorithm 1: Projected gradient descent for convex ERM

Choose an initial point $\theta_0 \in \Theta$, a fixed step size $\eta > 0$, and a number of iterations $T \geq 1$. For $t = 0, \dots, T - 1$, compute

$$g_t = \nabla_{\theta} \ell(\theta_t; X) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \ell(\theta_t; X_i), \quad \theta_{t+1} = \Pi_{\Theta}(\theta_t - \eta g_t).$$

Return the average of the points at which we computed gradients:

$$\hat{\theta} = \frac{1}{T} \sum_{t=0}^{T-1} \theta_t.$$

Theorem 10.1 (Convergence of projected gradient descent). *Under the assumptions above, let $\theta_X^* \in \arg \min_{\theta \in \Theta} \ell(\theta; X)$. For every fixed step size $\eta > 0$, the algorithm returns $\hat{\theta} \in \Theta$ satisfying*

$$\ell(\hat{\theta}; X) - \ell(\theta_X^*; X) \leq \frac{\|\theta_0 - \theta_X^*\|_2^2}{2\eta T} + \frac{\eta L^2}{2} \leq \frac{2R^2}{\eta T} + \frac{\eta L^2}{2}.$$

In particular, when $R, L > 0$, choosing $\eta = 2R/(L\sqrt{T})$ gives empirical excess risk at most $2RL/\sqrt{T}$.

Proof. Use the squared distance to the empirical minimizer as the potential:

$$\Phi_t = \|\theta_t - \theta_X^*\|_2^2.$$

Projection cannot increase distance to a point in Θ . Indeed, if $p = \Pi_\Theta(y)$, the optimality condition for projection gives $\langle y - p, \theta - p \rangle \leq 0$ for every $\theta \in \Theta$; expanding $\|y - \theta\|_2^2$ then gives $\|p - \theta\|_2^2 \leq \|y - \theta\|_2^2$. Applying this to the update and expanding the square,

$$\begin{aligned} \Phi_{t+1} &\leq \|\theta_t - \eta g_t - \theta_X^*\|_2^2 \\ &= \Phi_t - 2\eta \langle g_t, \theta_t - \theta_X^* \rangle + \eta^2 \|g_t\|_2^2 \\ &\leq \Phi_t - 2\eta (f(\theta_t) - f(\theta_X^*)) + \eta^2 L^2. \end{aligned}$$

The last inequality uses convexity and the gradient bound. Rearranging shows that each iteration's excess loss is paid for by a drop in potential, plus a step-size penalty:

$$f(\theta_t) - f(\theta_X^*) \leq \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{\eta L^2}{2}.$$

Summing over $t = 0, \dots, T-1$ telescopes the potential terms. Since $\Phi_T \geq 0$,

$$\frac{1}{T} \sum_{t=0}^{T-1} (f(\theta_t) - f(\theta_X^*)) \leq \frac{\Phi_0}{2\eta T} + \frac{\eta L^2}{2}.$$

Convexity of Θ gives $\hat{\theta} \in \Theta$, and convexity of f gives $f(\hat{\theta}) \leq \frac{1}{T} \sum_{t=0}^{T-1} f(\theta_t)$. Finally, $\|\theta_0 - \theta_X^*\|_2 \leq \|\theta_0\|_2 + \|\theta_X^*\|_2 \leq 2R$, proving the stated bound. Substituting $\eta = 2R/(L\sqrt{T})$ gives $2RL/\sqrt{T}$. $\square$

Using approximate gradients. Now suppose we cannot compute the empirical gradient exactly. Instead, at each iteration we obtain a vector g_t satisfying

$$\|g_t - \nabla f(\theta_t)\|_2 \leq \Delta.$$

We use the same update $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta g_t)$ and return the same average $\hat{\theta} = \frac{1}{T} \sum_{t=0}^{T-1} \theta_t$. The errors may be biased and may depend on all previous iterates; we only need the bound to hold at every iteration.

Theorem 10.2 (Convergence with approximate gradients). *Under the same assumptions on f and Θ , suppose $\|g_t - \nabla f(\theta_t)\|_2 \leq \Delta$ for every $t = 0, \dots, T-1$. For every fixed step size $\eta > 0$, the average iterate satisfies*

$$\ell(\hat{\theta}; X) - \ell(\theta_X^*; X) \leq \frac{\|\theta_0 - \theta_X^*\|_2^2}{2\eta T} + \frac{\eta(L + \Delta)^2}{2} + 2R\Delta.$$

In particular, when $R > 0$ and $L + \Delta > 0$, choosing $\eta = 2R/((L + \Delta)\sqrt{T})$ gives

$$\ell(\hat{\theta}; X) - \ell(\theta_X^*; X) \leq \frac{2R(L + \Delta)}{\sqrt{T}} + 2R\Delta.$$

Proof. Use the same potential $\Phi_t = \|\theta_t - \theta_X^*\|_2^2$. Convexity and Cauchy–Schwarz give

$$\begin{aligned} f(\theta_t) - f(\theta_X^*) &\leq \langle \nabla f(\theta_t), \theta_t - \theta_X^* \rangle \\ &= \langle g_t, \theta_t - \theta_X^* \rangle + \langle \nabla f(\theta_t) - g_t, \theta_t - \theta_X^* \rangle \\ &\leq \langle g_t, \theta_t - \theta_X^* \rangle + 2R\Delta. \end{aligned}$$

Also, $\|g_t\|_2 \leq \|\nabla f(\theta_t)\|_2 + \Delta \leq L + \Delta$. Expanding the projected update exactly as before,

$$\Phi_{t+1} \leq \Phi_t - 2\eta \langle g_t, \theta_t - \theta_X^* \rangle + \eta^2 (L + \Delta)^2.$$

Combining these inequalities and rearranging,

$$f(\theta_t) - f(\theta_X^*) \leq \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{\eta(L + \Delta)^2}{2} + 2R\Delta.$$

Sum over all T iterations, telescope the potential terms, and use convexity to pass to the average iterate. This proves the first bound. The second follows from $\Phi_0 \leq 4R^2$ and the stated choice of η . $\square$

Thus the optimization term decreases as $1/\sqrt{T}$, while the guarantee retains an additive $2R\Delta$ term from gradient error. This is the term we will control by using a robust mean estimator to approximate the gradient.